LA CONSERVACIÓN Y REUTILIZACIÓN DE LOS DATOS CIENTÍFICOS EN ESPAÑA. INFORME DEL GRUPO DE TRABAJO DE BUENAS PRÁCTICAS.

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

FECYT
FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA

Edición, Diseño y Maquetación
Fundación Española para la Ciencia y la Tecnología, FECYT, 2012

Conclusiones
Fundación Española para la Ciencia y la Tecnología, FECYT.

Autores
Grupo de Trabajo de "Depósito y Gestión de datos en Acceso Abierto" del proyecto RECOLECTA.

Coordinación
Cristina González Copeiro (FECYT)
Jordi Serrano-Muñoz (UPC)

Participantes
Alicia García-García (UCV)
Antonia Ferrer-Sapena (UPV)
Fernanda Peset (UPV)
Isabel Bernal (CSIC)
Izaskun Lacunza (FECYT)
Javier Gómez (UA)
Luís Martínez-Uribe (Fundación Juan March)
Manuela Palafox (UCM)
Mercedes de Miguel Estévez (FECYT)
Paz Fernández (Fundación Juan March)
Pilar Rico Castro (FECYT)
Ricard de la Vega (CESCA)
Victoria Rasero (UC3M)

Colaboradores
Agnes Ponsati (CSIC)
Florencia Dieci (UPV)

Fecha de edición
Diciembre 2012

Cómo citar este documento
Grupo de Trabajo de "Depósito y Gestión de datos en Acceso Abierto" del proyecto RECOLECTA. *La conservación y reutilización de los datos científicos en España. Informe del grupo de trabajo de buenas prácticas* [en línea] Madrid: Fundación Española para la Ciencia y la Tecnología, FECYT (2012) [Consulta 14/01/2013]. Disponible en WWW.FECYT.ES

Este informe está bajo una Licencia Creative Commons Atribución-NoComercial-SinDerivadas 3.0 Unported

SUMARIO

Introducción

Este informe surge para dar respuesta al reto que se abre dentro del movimiento de acceso abierto sobre cómo incluir los datos de investigación junto a las publicaciones científicas dentro de los repositorios. Contribuye de esta forma a la mejor aplicación de la Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación, en lo que se refiere al artículo 37 de difusión en abierto. Tiene por objetivo ayudar a la normalización de la gestión de los datos en los repositorios con el fin de facilitar su preservación, acceso y distribución. En su contenido se reflejan todos los aspectos importantes que intervienen en la gestión de los datos, desde su definición, tipos de datos, actores implicados, buenas prácticas para la gestión y un panorama general de la situación en España.

La Fundación Española para la Ciencia y la Tecnología (FECYT), en colaboración con Red de Bibliotecas Universitarias (REBIUN) de la Conferencia de Rectores de las Universidades Españolas (CRUE), gestiona y coordina RECOLECTA, un proyecto para la creación de una red de repositorios institucionales interoperables y que puede ser considerado como la primera iniciativa nacional en la creación de una infraestructura que facilita la *"open science"* o ciencia en abierto. El objetivo es además dotar de mayor visibilidad y servicios a los resultados de la investigación y de la producción científica española.

En el marco de este proyecto en 2012 se puso en marcha un grupo de trabajo cuyo objetivo fue el estudio del panorama general de la gestión de los datos científicos de investigación y su uso en el ámbito de los repositorios.

Nuestro agradecimiento a todas las instituciones participantes en el grupo de trabajo: la Universitat Politècnica de Catalunya (UPC), la Universidad Carlos III de Madrid (UC3M), la Universidad Complutense de Madrid (UCM), el Consejo Superior de Investigaciones Científicas (CSIC), la Universidad de Alicante (UA), el Centro de Servicios Científicos y Académicos de Cataluña (CESCA), el Instituto Juan March y la Universidad Politécnica de Valencia (UPV).

Confiamos en que este estudio resulte de ayuda e interés para la gestión de los datos de investigación.

1. LOS DATOS DE INVESTIGACIÓN

En los últimos años, el movimiento de Acceso Abierto a la información científica ha iniciado un debate sobre nuevas tendencias en el acceso, uso y modelos de negocio de la información producida con fondos públicos. Este movimiento tiene una presencia importante en el acceso abierto a publicaciones científicas publicadas en revistas. En este sentido, múltiples agencias de financiación e instituciones que realizan investigación ya disponen de políticas para garantizar el acceso abierto a publicaciones científicas financiadas con fondos públicos.

El movimiento de acceso abierto y de creación de e-infraestructuras que den apoyo al uso de información científica por la comunidad científica ha comenzado a debatir también sobre la importancia de los datos de investigación. Estos datos de investigación están empezando a reconocerse como una fuente de conocimiento propia e independiente de las publicaciones que pueden emplearse en la validación de los resultados de investigación publicados en artículos, para generar nuevo conocimiento y ser explotados por humanos y máquinas de manera interdisciplinar.

Para asegurar esta explotación de los datos, es necesario que estén disponibles y accesibles en la red, de la misma manera que lo están las publicaciones. Sin embargo, la naturaleza de los datos de investigación es muchísimo más variable y dependiente de la disciplina y de su particular ciclo de vida. Además, los requisitos técnicos y legales para garantizar el acceso son más complejos que los de las publicaciones. Existen ya disciplinas de la ciencia con tradición de depósito y re-uso de datos disponibles en repositorios temáticos, pero muchas otras que no han incluido esta práctica en sus rutinas de investigación. La gestión adecuada de los datos requiere, además, de inversión, personal especializado en la generación de datos, explotación de los mismos y su posterior preservación, coordinación para garantizar la interoperabilidad de los nodos de la infraestructura, cambio de cultura entre el personal investigador, etc.[1].

Actualmente, existe ya un acuerdo internacional para considerar la creación de una infraestructura transnacional y multidisciplinar que garantice el acceso a los datos de investigación, que contribuirá a mejorar la calidad de la ciencia, multiplicará sus resultados y evitará duplicidades[2] [3]. En este ámbito, se han dado múltiples pasos, particularmente desde agencias de financiación, para estimular la cultura "*open science*" que incluya a los datos de investigación como parte de una e-infraestructura que dé soporte a la ciencia del siglo XXI.

[1] A surfboard for riding the wave: Towards a four country action programme on research data; Knowledge Exchange, 2011; http://www.knowledge-exchange.info/Default.aspx?ID=469 [Consulta 6/12/2012]
[2] High level expert group on scientific data: Riding the Wave: How Europe can gain from the rising tide of scientific data; European Union, 2010; http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf [Consulta 6/12/2012]
[3] OECD Principles and Guidelines for Access to Research Data from Public Funding, OECD, 2007; http://www.oecd.org/dataoecd/9/61/38500813.pdf [Consulta 6/12/2012]

Sin tratar de ser exhaustivos, y a modo de muestra de las tendencias internacionales al respecto, se destacan en esta introducción algunos documentos y comunicaciones europeas que están marcando las tendencias en la redefinición del acceso a la información científica, concebida como una e-infraestructura a disposición de la comunidad investigadora y el público en general y disponible en acceso abierto cuando el conocimiento proceda de proyectos financiados con fondos públicos.

En 2007, la Comisión Europea publicó una comunicación sobre información científica en la era digital, destacando las primeras acciones previstas por la Comisión para coordinar el paso de la era de la información científica en papel al entorno digital[4]. Estas recomendaciones se centraban en facilitar el acceso a las publicaciones científicas, cofinanciar infraestructuras de investigación (repositorios), y estimular el debate para futuras políticas al respecto y el debate entre los diferentes actores.

A esta Comunicación, le siguieron las conclusiones del Consejo sobre información científica[5] que otorgan al acceso rápido a las publicaciones y datos de investigación el carácter de crucial para el desarrollo del Espacio Europeo de Investigación.

Fruto de estas conclusiones, la Comisión Europea lanzó un proyecto piloto en el ámbito del Séptimo Programa Marco, que estimulaba a los beneficiarios de siete áreas del programa a depositar sus artículos de investigación científica en repositorios temáticos o institucionales, respetando un periodo de embargo de entre 6 y 12 meses[6]. Como apoyo a este piloto, se financió también el proyecto OpenAire, que dotaba de infraestructura tecnológica y apoyo técnico para el cumplimiento del piloto[7].

También en 2007, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) publicó una guía para el acceso a los datos de información científica procedentes de financiación pública, que tenía por objetivo proveer de recomendaciones generales a los responsables de política científica y agencias de financiación de los estados miembros para estimular el acceso a los datos de investigación[8].

En 2010, la Comisión Europea encargó al "Grupo de alto nivel en datos de investigación" un informe con su visión sobre el acceso, uso, re-uso y calidad de los datos de investigación científica en 2030[9]. Este informe

[4] Communication on scientific information in the digital age: access, dissemination and preservation (Com 2007 56 Final); http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf [Consulta 6/12/2012]
[5] Council Conclusions on scientific information in the digital age: access, dissemintation and preservation, European Union, 2007; http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/intm/97236.pdf [Consulta 6/12/2012]
[6] Commission Decision on the adoption and a modification of special clauses applicable to the model grant agreement of FP7 C(2008) 4408 final http://ec.europa.eu/research/press/2008/pdf/decision_grant_agreement.pdf [Consulta 6/12/2012]
[7] OpenAire FP7 project http://www.openaire.eu/ [Consulta 6/12/2012]
[8] OECD Principles and Guidelines for Access to Research Data from Public Funding, OECD, 2007; http://www.oecd.org/dataoecd/9/61/38500813.pdf [Consulta 6/12/2012]
[9] High level expert group on scientific data: Riding the Wave: How Europe can gain from the rising tide of scientific data; European Union, 2010; http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf [Consulta 6/12/2012]

está sirviendo como hoja de ruta europea en la consecución de una e-infraestructura que maximice los beneficios del acceso a la información científica.

En respuesta a este informe, "*Knowledge Exchange"*, una asociación con miembros de instituciones dedicadas a la creación de e-infraestructuras para la investigación y la enseñanza superior de cuatro países europeos, ha elaborado una propuesta para la creación de un plan de acción de Gran Bretaña, Dinamarca, Holanda y Alemania sobre datos de investigación[10].

La Comisión Europea prepara para finales del año 2012 unas nuevas recomendaciones sobre acceso abierto y preservación de la información científica, que previsiblemente ahondarán en el estímulo del contenido científico en abierto (de publicaciones y datos), las infraestructuras abiertas e interoperables y la "*open culture*" (para investigadores y público en general).

A nivel nacional, la recién aprobada "Ley de la ciencia, la tecnología y la innovación"[11] se suma al estímulo de creación de infraestructuras de apoyo a la información científica, con un artículo especialmente dedicado al depósito en repositorios institucionales o temáticos de artículos científicos financiados con Presupuestos Generales del Estado.

El presente informe surge del ámbito del proyecto Recolecta y pone de manifiesto algunas consideraciones importantes que han de tenerse en cuenta en el diseño e implementación de una política de gestión de datos de investigación, con especial énfasis en la situación de España con respecto a otros países. A lo largo de este informe, se define la variedad de tipos de datos de investigación, los actores implicados en su gestión (los repositorios institucionales y temáticos, las agencias de financiación, los centros de datos existentes, investigadores, bibliotecarios y expertos en la gestión de datos, etc.). Asimismo, se reflexiona sobre los aspectos económicos derivados de la creación de una infraestructura interoperable de gestión de datos. Por último, el informe pretende contribuir a futuras iniciativas que van a ser necesarias adoptar para la gestión de los datos resultado de la investigación, en el ámbito de la nueva Ley de la Ciencia, Tecnología e Innovación.

[10] A surfboard for riding the wave: Towards a four country action programme on research data; Knowledge Exchange, 2011; http://www.knowledge-exchange.info/Default.aspx?ID=469 [Consulta 6/12/2012]
[11] Ley 14/2011 de la Ciencia, la Tecnología y la Innovación http://www.boe.es/boe/dias/2011/06/02/pdfs/BOE-A-2011-9617.pdf [Consulta 6/12/2012]

2. ACTORES IMPLICADOS EN LA GESTIÓN DE LOS DATOS CIENTÍFICOS

La e-ciencia ha cambiado las prácticas de la investigación en todas las áreas científicas. El aumento de la capacidad computacional permite a los investigadores procesar y compartir grandes cantidades de información. Para facilitar la reutilización de los datos científicos hay que adoptar los estándares utilizados por la comunidad de datos de investigación, desarrollar y promocionar guías de buenas prácticas que ayuden a los investigadores a gestionar adecuadamente sus datos de investigación, impulsar programas de formación que doten a la comunidad científica de las competencias necesarias, proteger la propiedad intelectual de los productores de datos y establecer los mecanismos necesarios para asegurar la calidad. Para ello, es fundamental alcanzar un alto grado de coordinación entre los agentes implicados en la gestión de los datos.

En este apartado se describe el papel que desempeñan los actores que intervienen en la gestión de los datos científicos y las responsabilidades asociadas[12].

- ## Investigadores/productores de datos

Proporcionan la evidencia y validación científica de las investigaciones. Si bien esta categoría se compone fundamentalmente de investigadores, en algunos casos hay conjuntos de datos que ya existen y los científicos los utilizan para validar sus tesis. La comunidad investigadora puede ser considerada como productores, autores, y usuarios de los datos de investigación.

- ## Universidades y Centros de Investigación

Su principal responsabilidad es establecer la política interna de gestión de los datos científicos. Establecen los estándares para los distintos tipos de datos y la guía de buenas prácticas. Las instituciones deben asumir la responsabilidad de promoción para que los resultados de la investigación de sus investigadores se depositen en los repositorios institucionales para su custodia y preservación a corto plazo, proporcionando la formación adecuada.

Dentro de las Universidades y centros de Investigación, cabe destacar aquellos servicios más relevantes en la gestión de datos como los servicios de Informática, Bibliotecas y Servicios de Investigación. Cada uno de ellos tiene papeles complementarios (el de Informática en almacenamiento; el de Biblioteca en metadatos, apoyo a la publicación y derechos; y los Servicios de Investigación en políticas institucionales, planes de gestión y temas de ética) y es necesario que se coordinen para poder dar un servicio institucional completo.

- ## Repositorios institucionales

Desempeñan un papel básico en el almacenamiento de los datos a corto plazo, frente al papel que tienen los centros de datos de preservación a largo plazo. Es básico el uso de estándares que facilite la interoperabilidad entre los repositorios y los centros de datos. Es muy importante la fiabilidad y robustez de los enlaces y el establecimiento de mecanismos para la migración de datos entre los repositorios, así

[12] Lyon, Liz (2007) Dealing with Data: Roles, Rights, Responsabilities and Relationships. Consultancy Report. UKOLN http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.doc [Consulta 6/12/2012]

como el mantenimiento de las versiones de los datos si se encuentran en distintos espacios de almacenamiento. La sostenibilidad del archivo de los datos de investigación constituye uno de los desafíos y problemas clave.

- **Centros de datos**

Establecen guías de buenas prácticas y la selección de los datos que deben preservarse a largo plazo, facilitando su difusión. Protegen los derechos de propiedad de los productores de los datos y proporcionan herramientas para su reutilización. Desarrollan planes de recuperación de datos en caso de desastres.

- **Gestores de datos**

El perfil profesional del gestor de datos requiere competencias informáticas, conocimiento de la disciplina, de las prácticas de investigación y flujos de trabajo, comprensión de las normas técnicas específicas, esquemas de metadatos y vocabularios de uso habitual.

También deben conocer cuáles son los centros de datos nacionales e internacionales de la investigación en la disciplina y disponer de un buen conocimiento de los requisitos de publicación de datos de las revistas académicas más importantes[13]. La responsabilidad de los gestores de datos es gestionar y promocionar el uso de datos desde su creación para asegurar su uso y su disponibilidad para ser localizados y reutilizados[14].

- **Usuarios que reutilizan los datos**

Deben cumplir las condiciones de la licencia y los permisos de utilización, reconociendo los derechos de propiedad intelectual de los investigadores productores de los datos.

- **Agencias de financiación**

Las agencias de financiación implementan las políticas de datos con los actores implicados, determinan las fechas de preservación, resuelven problemas de confidencialidad, protección de datos y uso de licencias. Desde comienzos de 2000, las agencias de financiación de algunos países (*National Institutes of Health, Wellcome Trust*, etc.) han comenzado a pedir la liberación de datos en diversos grados y con diferentes niveles de cumplimiento, con el fin de maximizar el retorno de la financiación a la investigación. Desde 2010, la *National Science Foundation* exige que las propuestas de financiación vayan acompañadas de un Plan de Gestión de Datos[15].

- **Publicaciones científicas**

Del mismo modo que las agencias de financiación, los editores de publicaciones científicas están enlazando los artículos de las revistas con los datos de investigación utilizados, con el fin de compartir esos datos con lectores e investigadores.

[13] Lyon, Liz (2012) The Informatics Transform: Re-Engineering Libraries for the Data Decade. The International Journal of Digital Curation. Volume 7, Issue 1, 2012
http://www.ijdc.net/index.php/ijdc/article/view/210/279 [Consulta 6/12/2012]

[14] Martínez-Uribe, Luis, Macdonald, Stuart (2008). Un nuevo cometido para los bibliotecarios académicos: data curation. El profesional de la información, v.17, n. 3, mayo-junio 2008

[15] Borgman, C.L. (2011). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=186915 [Consulta 6/12/2012]

3. ¿QUÉ SON LOS DATOS DE LA INVESTIGACIÓN?

3.1 Definición

Definir los datos de la investigación no es tarea sencilla, los datos producidos por los investigadores forman un grupo de materiales extremadamente heterogéneo y complejo, creado para distintos propósitos y mediante procesos también diferentes. Los datos son el *"alma"* de la investigación, rara vez son objetos sencillos que pueden ser fácilmente compartidos, sino que encarnan las perspectivas epistemológicas de sus creadores[16].

La Universidad Australiana de Melbourne aporta la siguiente definición en su política institucional de datos:

Los datos de la investigación son hechos, observaciones o experiencias en que se basa el argumento, la teoría o la prueba. Los datos pueden ser numéricos, descriptivos o visuales. Los datos pueden ser en estado bruto o analizado, pueden ser experimentales u observacionales. Los datos incluyen: cuadernos de laboratorio, cuadernos de campo, datos de investigación primaria (incluidos los datos en papel o en soporte informático), cuestionarios, cintas de audio, videos, desarrollo de modelos, fotografías, películas, y las comprobaciones y las respuestas de la prueba. Las colecciones datos para la investigación pueden incluir diapositivas; diseños y muestras. En la información sobre la procedencia de los datos también se podría incluir: el cómo, cuándo, donde se recogió y con que (por ejemplo, instrumentos). El código de software utilizado para generar, comentar o analizar los datos también pueden ser considerados datos.

3.2 Tipos de datos

La National Science Foundation (2007) propone la siguiente categorización de datos de investigación basada en su origen que ayuda a comprender mejor la variedad de tipos y sus distintas necesidades a la hora de gestionarse:

- **Datos observacionales.** Son registros históricos, se pueden obtener únicamente en un lugar y en un momento en el tiempo. Esta característica los hace especialmente importantes a la hora de preservarlos ya que en caso de que se perdiesen no podrían volver a reproducirse. Ejemplos: los barómetros del Centro de Investigaciones Sociológicas (CIS), son encuestas de opinión sobre diversos temas que preocupan a los españoles. El Banco Nacional de Datos Climatológicos sería otro caso de este tipo ya que posee información sobre precipitaciones registradas en España desde hace 150 años.

- **Datos experimentales.** Son los datos que acompañan a los experimentos desde su planificación y preparación hasta la obtención de resultados. Los experimentos en muchos casos pueden repetirse

[16] Borgman, CL (2012) On Local or Global? Making Sense of the Data Sharing Imperative. Talk at University of Southern Carolina on 9th April 2012

para obtener los mismos datos sin embargo en ocasiones el coste de repetir el experimento hace que no sea rentable repetirlo. Ejemplos: el acelerador de partículas del CERN en Ginebra produce una cantidad desorbitada de datos experimentales capaz de llenar 100,000 DVDs al año. En los laboratorios de investigación ya sean químicos, biológicos o en otras disciplinas también se producen gran cantidad de datos con instrumentos especializados.

- **Datos computacionales.** Estos son los datos que acompañan a las simulaciones que suelen incluir datos de entrada, ciertos programas y resultados. Para este tipo de datos en la mayoría de los casos no se necesitan los resultados ya que con los datos de entrada, los programas y el ordenador que los genera debiera de ser posible reproducirlos. Ejemplos: pueden ser datos producidos en centros de computación avanzada que simulan el funcionamiento de órganos del cuerpo humano, el movimiento de los astros o predicen el tiempo.

De esta manera cada disciplina científica basará su investigación en estas tipologías y en aquellas en las que se puedan subdividir. Ya sean cualitativos, cuantitativos, geográficos, espaciales, u otros, pertenecerán a uno o a varios de los ejes mencionados.

3.3 La gestión de los datos

La correcta gestión de los datos de investigación es una parte fundamental de proceso de investigación. Esta gestión consiste en la toma de decisiones y acciones desde antes de la creación de los datos, durante su creación y uso y a lo largo de su ciclo de vida. Algunas de las etapas que debe de incluir una correcta gestión de datos son:

- Un plan de gestión de datos como parte de la propuesta de financiación que anticipe los retos de la gestión y proponga soluciones a los mismos.
- Tratar las cuestiones éticas y legales oportunas referentes a datos personales sensibles, copyright y licencias de acceso y uso de los datos.
- La organización y documentación de los datos de acuerdo a estándares disciplinares e internacionales que permitan conocer qué son los datos y como se crearon los datos para poder ser reutilizados.
- Mecanismos apropiados de almacenamiento, back-up y seguridad de la información que aseguren la confidencialidad, integridad y disponibilidad de la información.
- Compartir los datos de manera que se citen de forma estándar y así dar crédito a los creadores de los mismos.
- Archivo de una copia final de los datos en centros de datos especializados que tomen las medidas necesarias para la preservación y difusión de los datos.

Para que sea posible gestionar los datos de este modo es necesario que existan políticas, a nivel de agencias de financiación e institucional, que definan y aclaren los papeles y responsabilidades de los distintos actores. La responsabilidad de esta gestión a lo largo del ciclo de vida debe recaer en una variedad de instituciones tales como las agencias de financiación, las Universidades, las Bibliotecas, los Centros Informáticos y los propios investigadores. Pero ante todo han de ser los investigadores y sus necesidades el punto de partida.

La Ligue des Bibliothèques Européennes de Recherche - Association of European Research Libraries (LIBER) creó en 2010 un grupo de trabajo sobre e-Ciencia (Working Group on e-Science), el resultado ha sido un informe final[17] que incluye diez recomendaciones para las bibliotecas que se inicien en la gestión de datos de investigación, en las conclusiones se destaca que las bibliotecas pueden y deben desempeñar tareas en el apoyo a los investigadores en la gestión y planificación de los datos.

[17] Christensen-Dalsgaard, Birte et al (2012) Ten recommendations for libraries to get started with research data management: Final report of the LIBER working group on E-Science / Research Data Management. http://www.libereurope.eu/sites/default/files/WGSC_20120801.pdf [Consulta 9/12/2012]

4. INFRAESTRUCTURA Y SOSTENIBILIDAD

Los datos han de ser gestionados por una infraestructura fiable y estable que asegure la confiabilidad y su integridad. El white paper "Strategy for a European Data Infrastructure"[18] recoge los principales requisitos de infraestructura de diversas iniciativas de datos de algunas disciplinas y comunidades de investigación a nivel europeo. En resumen son:

- Preservación de datos a largo plazo incluyendo mecanismos de autenticidad y de control de calidad de los datos.
- Acceso a los datos (ciclo de vida de los datos), servicios de data curation y capacidad de computación en la infraestructura (data mining, data processing...).
- Distribución de los datos y federaciones, no solo por motivos de preservación sino también para la optimización y aumento del rendimiento del acceso.

A estos requisitos se les suma que los datos deben estar duplicados para conseguir la alta disponibilidad, requisito común de este tipo de sistemas.

Tres aspectos se han de tener en cuenta para dar solución a estos requisitos:

- Sistemas software capaces de gestionar el ciclo de vida de los datos.
- Sistemas de almacenamiento masivo de datos. Existen diversas tecnologías para este propósito, como la arquitectura NAS (Network Attached Storage) de crecimiento horizontal, que permite escalar rápidamente mediante nodos de tipo "comodity" en función de la demanda. En relación al ciclo de vida de los datos, pueden existir muchos factores dependientes de su naturaleza o disciplina, sin embargo, a nivel de flujos de bits que se almacenan en un soporte físico, pueden ser tratados de manera homogénea.
- Redes de alta capacidad para la transmisión de datos entre diferentes nodos. En España la red académica y de investigación española (RedIRIS) proporciona estos servicios avanzados de comunicaciones a la comunidad científica y universitaria nacional.

Estas infraestructuras han de ser tenidas en cuenta de cara a analizar la viabilidad de las iniciativas para la gestión de datos, pues sus costes, tanto de adquisición como de mantenimiento, son elevados. Se estima que los costes de mantenimiento de repositorios de datos científicos son de un orden de magnitud mayor que los tradicionales repositorios de publicaciones[19].

[18] Strategy for a European Data Infrastructure http://www.csc.fi/english/pages/parade [Consulta 6/12/2012]

[19] Beagrie N, Chruszcz J and Lavoie B (2008). Keeping Research Data Safe 1. JISC http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf [Consulta 12/12/12]

Existen dos principios básicos para rentabilizar mejor estos costes:

- Procesos de selección de los datos. No todos los datos han/pueden ser enriquecidos (data curation) o preservados. Una buena selección integrada dentro del ciclo de vida de los datos y realizada desde el punto de vista del conocimiento específico de los datos y pensando no sólo en su uso principal, sino también en cómo podrán estos datos ser re-usados a posteriori es esencial.
- Uso de las economías de escala con respecto a las infraestructuras. Se trata de conseguir una capa de datos que agrupe infraestructuras de manera transversal, tal y como se realiza por Geant, RedIRIS o la Anella Científica en la capa de conectividad, o como los proyectos Driver hacen interoperables distintos repositorios de investigación. No sólo se conseguiría la compartición de costes, sino que además se aumentarían las sinergias entre distintos grupos de investigación o incluso entre diferentes disciplinas.

En apartados anteriores se menciona que los datos pueden ser muy heterogéneos, y dependiendo de ellos los costes asociados a la infraestructura pueden variar sustancialmente. En un extremo de altos costes en infraestructura se situarían proyectos con masivos datasets como los de los datos producidos por el Large Hadron Collider o el European Bioinformatics Institute, mientras que en el otro extremo, por ejemplo, se situaría el Worldwide Protein Data Bank Archive, repositorio con más de 80.000 estructuras en 3D de moléculas, pero que apenas ocupan 150GB de almacenamiento. En este último caso, los costes de infraestructura no son significativos comparados con las 69 FTE de personal que trabaja en el proyecto[20].

Aún sólo gestionando aquellos datos que sean "útiles" o imprescindibles, haciéndolo en infraestructuras que aprovechen las economías de escala y sea cual sea el tamaño de la infraestructura necesaria, para la gestión de datos científicos son necesarias políticas de financiación de las infraestructuras a largo plazo, pues los datos son acumulativos y se preservan típicamente más allá de los ciclos tecnológicos.

Como se ha mencionado antes, en las propuestas de financiación de proyectos, se debería detallar un plan de gestión de datos, incluyendo su viabilidad económica.

[20] The Royal Society (2012). Science as an open enterprise http://royalsociety.org/policy/projects/science-public-enterprise/report [Consulta 12/12/12]

5. BUENAS PRÁCTICAS PARA LA GESTIÓN DE DATOS DE INVESTIGACIÓN

Los datos de investigación constituyen uno de los principales activos en el proceso de investigación científica. Una óptima gestión de dichos datos favorece la innovación y el desarrollo de la misma, puesto que permitiría la explotación de datos de alta calidad (compartir – reutilizar).

En el marco global de la E-Ciencia, el objeto específico del control, organización, descripción y preservación de datos científicos es el 'dataset', que se define como una colección de datos reunidos durante la ejecución de un proyecto de investigación. Los datasets son objetos digitales compuestos y heterogéneos. Es decir, pueden comprender diferentes elementos o tipos de datos: documentos de texto, hojas de cálculo, ficheros de operaciones matemáticas, gráficos, imágenes, etc. El dataset constituye la base de una investigación y va asociado a una publicación científica como resultado de dicha investigación. El dataset adquiere valor añadido si se integra con la publicación relacionada ('linking data': cita y enlace), independientemente de su ubicación.

Los datasets se almacenan y gestionan en repositorios interoperables en red integrados en una infraestructura global de investigación, desarrollados conforme a estándares internacionales.

Instituciones de educación superior y agencias de financiación de la investigación de varios países están llevando a cabo iniciativas para crear infraestructuras de gestión de datos que posibiliten la reutilización de los datasets, mediante la adopción de políticas que promueven el acceso abierto y la compartición de los datos, y garantizando la sostenibilidad y accesibilidad de los datos a largo plazo.

El movimiento Open Data, en el marco del Open Access, define los datos abiertos como aquéllos que se pueden usar, reutilizar y redistribuir sin otra restricción que el requisito de atribución o compartir igual[21].

5.1 Desarrollo de un plan de gestión de datos

La responsabilidad de la gestión de los datos corresponde en primer lugar a los investigadores, pero las instituciones deben proporcionar el soporte técnico y organizativo a su comunidad. Organizativamente, en un servicio de gestión de datos de investigación, es imprescindible la colaboración entre los investigadores y productores de los datos y los bibliotecarios de datos dentro de una institución.

Los investigadores son los expertos que deben proporcionar la información contextual necesaria para determinar el origen y el ciclo de vida de los datos. Los bibliotecarios son expertos en la gestión de información y han de proporcionar apoyo especializado y personalizado a los investigadores, así como

[21] http://opendefinition.org/okd/ [Consulta 6/12/2012]

utilizar los medios técnicos necesarios para que los datos sean comprendidos e interpretados por otros investigadores.

Dada la diversidad de datos científicos, por su naturaleza heterogénea y por la cultura específica de cada comunidad científica, la institución debe proporcionar a los investigadores un modelo de plan de gestión de los datos para ahorrar tiempo y esfuerzo en el proceso de la investigación[22]. La planificación conlleva una serie de ventajas:

- Se pueden encontrar y comprender los datos cuando se necesite utilizarlos.
- Se garantiza la continuidad del proyecto independientemente de la participación de los investigadores.
- Se evitan duplicaciones y tareas innecesarias.
- El mantenimiento del conjunto de datos generados permite la validación de los resultados.
- Los datos se pueden compartir permitiendo un alto nivel de colaboración y de avance en la investigación.
- Si los datos se ofrecen en abierto tendrán una gran visibilidad.
- Otros investigadores que utilicen los datos pueden citarlos y la investigación obtendrá más prestigio.

La descripción mínima de los datos debe tratar los siguientes aspectos:

- Contexto, descripción del proyecto y propósito de la investigación, metodología utilizada;
- Naturaleza de los datos, historia de los datos, contenido y estructura, terminología, software, fecha de creación y fechas de modificación, versiones, responsables y participantes;
- Formatos de ficheros, estructura y nomenclatura de los ficheros, sistema de almacenamiento, procedimiento para copias de seguridad;
- Aspectos legales, políticas de acceso y seguridad;

El paradigma tecnológico de un sistema de gestión de datos científicos incluye los siguientes requerimientos:

- El modelo lógico de datos (relacional) y su sistema de gestión (base de datos) han de permitir su descripción, su representación y su recuperación;
- El sistema de gestión deberá permitir una óptima organización de los datos, documentarlos, preservarlos y hacerlos accesibles;

[22] Existen herramientas para la elaboración de planes de este tipo, como por ejemplo DMPTool (https://dmp.cdlib.org/ [Consulta 6/12/2012])

- Un software que sea capaz de analizar gran cantidad de datos, procesarlos, tratarlos y obtener diferentes productos secundarios ('Data Mining').

5.2 Formatos

El formato en el que se archivan los datos es un factor primordial para asegurar su preservación y su accesibilidad. La evolución de las tecnologías son la causa de que tanto el hardware como el software se vuelvan obsoletos. Los investigadores utilizarían el formato y software adecuado a sus necesidades, pero para garantizar el acceso y la preservación a largo plazo, habría que tener en cuenta las siguientes consideraciones:

- Deben utilizarse, en la medida de lo posible, formatos abiertos, no propietarios.
- El formato utilizado ha de permitir la indización del contenido para su potencial recuperación.
- Un formato de compresión de datos utiliza menos espacio de almacenamiento.
- El formato elegido deberá ser estándar (IANA mime types), o estándar de facto para la comunidad investigadora.

Los ficheros y las carpetas deben estar bien organizados con una estructura ordenada. El sistema de nomenclatura es importante para identificar los contenidos.

Es necesario llevar un control de versiones de los ficheros para que puedan localizarse las sucesivas versiones y pueda conocerse los cambios de una con respecto a otra.

5.3 Metadatos

Los metadatos son un conjunto de información estructurada que ha de recoger el origen, propósito, referencia temporal, localización geográfica, creador, condiciones de acceso y términos de uso de un dataset. Los metadatos cumplen diferentes funciones relacionadas entre sí: la gestión y administración, la preservación, la descripción, la diseminación de los datos y la recuperación de los datos. La documentación y descripción de los datos facilita su localización, su comprensión y su utilización.

La documentación del dataset facilitada por el investigador se incluirá en el registro de metadatos. Los metadatos deben incluir al menos la siguiente información:

- *Título*: Nombre del proyecto del conjunto de datos o de investigación que lo produjo
- *Nombres de los creadores* y las *direcciones* de la organización o personas que han creado los datos.
- *Código de identificación de los datos*, incluso si es una referencia de uso interno.
- Palabras o frases que describen el *tema o el contenido de los datos.*
- *Patrocinadores*: Las organizaciones o agencias que financiaron la investigación.
- *Derechos*: Cualquier tipo de derechos de propiedad intelectual de los datos.
- *Acceso a la información*: ¿Dónde y cómo sus datos pueden ser accesibles por otros investigadores?

- *Idioma del contenido.*
- *Fechas clave asociadas a los datos*, incluyendo: inicio del proyecto y la fecha de finalización, fecha de lanzamiento, período de tiempo cubierto por los datos, y otras fechas relacionadas con la vida útil de datos, por ejemplo, el ciclo de mantenimiento, actualización del programa.
- *Lugar al que hacen referencia los datos* (p.e, una ubicación física, cobertura espacial etc).
- *Metodología:* ¿Cómo se generaron los datos, incluidos los equipos o el software utilizado, el protocolo experimental, etc?
- *Procesamiento de datos*: toda la información acerca de cómo los datos se han alterado o procesado.
- *Fuentes*: Citas a los materiales para los datos procedentes de otras fuentes, incluidos los detalles de los datos de origen.
- *Lista de nombres de archivo de la lista de todos los archivos de datos asociados con el proyecto*, con sus nombres y extensiones de archivo (por ejemplo, 'stone.mov').
- *Formatos de archivo de los datos*, por ejemplo, FITS, SPSS, HTML, JPEG, RIF-CS y el software necesario para leer los datos.
- *Organización de archivos*: estructura del archivo de datos (s) y la disposición de las variables, cuando sea aplicable.
- *Lista de variables* en los archivos de datos.
- *Explicación de los códigos o abreviaturas utilizadas* en cualquiera de los nombres de los archivos o las variables en los archivos de datos.
- *Versiones* de fecha / fecha y hora para cada archivo, y usar un ID diferente para cada versión (ver la organización de sus archivos).
- Operaciones de comprobación para verificar si los archivos han cambiado a lo largo del tiempo. (Algoritmo Checksum para proteger la integridad de los datos).

Los metadatos se estructuran en registros conforme a esquemas normalizados. Los criterios de adopción de un esquema u otro dependerán de los objetivos que se plantee la organización para la gestión de los datos. Para lograr la interoperabilidad con otros sistemas de gestión de datos es prioritaria la normalización. Con objeto de cumplir todas las funciones antes mencionadas, se suelen combinar diferentes esquemas de metadatos mediante la declaración del espacio de nombres correspondiente a cada esquema.

Existen varios estándares de metadatos, aunque aquí citaremos aquéllos cuyo uso está más extendido:

- ***Dublin Core Metadata Terms***[23]. Es un esquema muy simple de carácter universal, que puede ser aplicado a recursos de todo tipo y procedencia.

[23] http://dublincore.org/documents/dcmi-terms/ [Consulta 6/12/2012]

- ***Data Documentation Initiative (DDI)***[24]. Es un esquema diseñado específicamente para la descripción de conjuntos de datos de índole social y económico. Permite documentar el ciclo de vida completo de los datos.
- ***General International Standard Archival Description (ISAD(G))***[25]. Es un conjunto de elementos para describir archivos con varios niveles de agregación. Los procesos descriptivos pueden ser simultáneos a la producción de los documentos y continuar a lo largo de todo su ciclo vital.
- ***Metadata Encoding and Transmission Standard (METS)***[26]. Se trata de una norma para la codificación y agrupación de metadatos administrativos, técnicos, de preservación y descriptivos, que permite la representación de objetos digitales complejos con gran exhaustividad. También permite expresar las relaciones entre las partes de un objeto digital, así como las relaciones entre distintos objetos.
- ***ISO 19115 for geographic information***[27]. Esquema utilizado para la descripción de información y servicios geográficos. Es aplicable a los datasets geográficos.

Los registros de metadatos se agrupan en sistemas de búsqueda y recuperación de información, y podrán ser recolectados a través del protocolo OAI-PMH.

5.4 Identificador digital de datos

El dataset almacenado debe asociarse a un identificador digital único y persistente que facilite la verificación de los datos, la reutilización, la diseminación y el impacto, y el acceso a largo plazo. Los identificadores conforme a los preceptos de la web semántica deben tener forma de URI. La URI es una cadena de caracteres que condensa la dirección URL (Uniform Resource Location) y el nombre URN (Uniform Resource Name) del recurso[28].

Hay muchos sistemas diferentes, como por ejemplo:

- *PURL Uniform Resource Locator*. Funcionalmente, un PURL es una URL. Sin embargo, en lugar de apuntar directamente a la ubicación de un recurso de Internet, algunos puntos PURL apuntan a un servicio de resolución intermedia. El servicio de resolución de PURL asocia el PURL con la dirección URL real y devuelve la URL para el cliente.
- *DOI Digital Object Identifier*. Es un nombre para una entidad en las redes digitales. Proporciona un sistema de identificación permanente y viable y el intercambio interoperable de la información manejada en las redes digitales.

[24] http://www.ddialliance.org/what [Consulta 6/12/2012]
[25] http://www.icacds.org.uk/eng/ISAD(G)es.pdf [Consulta 6/12/2012]
[26] http://www.loc.gov/standards/mets [Consulta 6/12/2012]
[27] http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020 [Consulta 6/12/2012]
[28] http://www.w3.org/TR/uri-clarification/ [Consulta 6/12/2012]

- *ACCESSION*[29] – Números usados por el National Center for Biotechnology Information (NCBI) son únicos y citables.
- *InChI*[30] La IUPAC International Chemical Identifier (InChITM) es un identificador no propietario de las sustancias químicas que pueden ser utilizados en las fuentes de datos impresos y electrónicos, permitiendo así una vinculación más fácil de las compilaciones de datos diversos.

5.5 Marco legal relacionado con la gestión y divulgación de datos de investigación

La producción, gestión y diseminación de datos debe ajustarse a un marco legal en el que existen derechos y acuerdos que deben ser respetados. Las cuestiones clave al respecto serían:

- ¿Qué derechos legales existen sobre los datos y datasets?
- ¿A quién pertenecen estos derechos?
- ¿Qué restricciones legales se han de aplicar para la diseminación de los datos y datasets?
- ¿Qué contratos, permisos y licencias hay que utilizar para cumplir con la legalidad vigente?

Hay que tener en consideración los siguientes derechos:

- Los derechos de propiedad intelectual
- La confidencialidad, privacidad y protección de datos

Acceso y datos: Teniendo en cuenta las restricciones legales, es necesario identificar qué datos serán accesibles, identificar quién puede acceder a los datos y con qué propósito. Según la naturaleza de los datos debemos atender a las siguientes categorías:

- Datos públicos: pueden ponerse sin restricciones a disposición de cualquier usuario en acceso abierto.
- Datos restringidos: sólo pueden ser consultados por determinados usuarios.
- Datos privados: no se pueden hacer públicos. Son confidenciales.

Privacidad y confidencialidad: Cualquier investigación que contenga datos de carácter personal tiene que cumplir los preceptos de la legislación de protección de datos. En España la norma que regula estos aspectos es la *Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal,* cuyo objeto es "garantizar y proteger en lo que concierne al tratamiento de los datos personales, las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su honor e intimidad personal y familiar". La ley es de aplicación a los datos de carácter personal registrados en cualquier soporte físico. El tratamiento de los datos cubre las actividades de recolección, registro, almacenamiento,

[29] http://www.ncbi.nlm.nih.gov/ [Consulta 6/12/2012]
[30] http://www.iupac.org/home/publications/e-resources/inchi.html [Consulta 6/12/2012]

recuperación, consulta, uso y diseminación. Para garantizar el derecho a la protección de datos, es necesario informar a las personas implicadas y solicitar su consentimiento para el tratamiento de sus datos.

Propiedad intelectual y datos: En España la norma principal que regula los derechos de propiedad intelectual es la Ley de Propiedad Intelectual (*Real Decreto Legislativo 1/1996 de 12 de abril por el que se aprueba el Texto Refundido de la LPI*) que ha sufrido varias modificaciones, entre ellas la operada por la Ley 23/2006 de 7 de julio con el objeto de adaptar la normativa española a las nuevas circunstancias creadas por la sociedad de la información.

Las colecciones de datos y las bases de datos están protegidas por propiedad intelectual, según el art. 12 del mencionado TRLPI mediante el denominado derecho *sui generis*, en cuanto que constituyen creaciones intelectuales. "La protección se refiere únicamente a su estructura en cuanto forma de expresión de la selección o disposición de contenidos", no a los datos mismos. Los derechos de autor pertenecen a sus creadores, siempre que se trate de trabajos originales.

Los derechos morales son derechos de carácter personal que pertenecen exclusivamente a los autores y son irrenunciables. En virtud de estos derechos corresponde a los autores fundamentalmente, el decidir si su obra ha de ser divulgada y en qué forma, y exigir el reconocimiento de la autoría.

Los derechos de explotación o *copyrigh*t son transferibles. El titular de estos derechos posee su ejercicio exclusivo y no pueden ser realizados sin su autorización, salvo en los límites que establece la ley. Los derechos de explotación constituyen una serie de actos como el de reproducción, distribución, comunicación pública y transformación.

Existen excepciones al ejercicio de los actos de explotación, como en el caso de reproducción para uso exclusivamente privado, usos en beneficio de personas con discapacidad, uso a título de cita o ilustración con fines educativos.

Las obras en situación de dominio público, cuando el plazo de protección de los derechos ha expirado, pueden ser utilizadas de forma libre y gratuita[31].

Depósito de los datos: El depósito de los datasets en un repositorio implica el ejercicio de los derechos de explotación, por lo que se requiere el permiso explícito del titular de dichos derechos mediante un acuerdo de cesión no exclusiva de los derechos necesarios.

[31] El TRLPI establece un plazo de duración de los derechos de una obra en setenta años desde su divulgación, y en setenta años desde su creación si no han sido divulgadas.

Conforme al movimiento "Open access", los datos resultantes de proyectos financiados con fondos públicos constituyen un bien de interés público, por lo que deben estar disponibles en un repositorio en acceso abierto sin perjuicio de preceptos legales o éticos.

Licencias alternativas al copyright: Como hemos mencionado anteriormente, el titular de los derechos de explotación tiene la potestad de determinar quién puede acceder a los datos y bajo qué condiciones. Existen licencias estándares y libres que el autor puede aplicar a sus datos de investigación para proporcionar los términos en los que compartir y reutilizar dichos datos en el ámbito de Internet. Un ejemplo de dichas licencias son las Creative Commons, que en conjunto constituyen seis licencias que permiten la copia, distribución, descarga y transformación de los documentos digitales:

	**RECONOCIMIENTO (*Attribution*):** En cualquier explotación de la obra autorizada por la licencia hará falta reconocer la autoría.
	**NO COMERCIAL (*Non Commercial*):** La explotación de la obra queda limitada a usos no comerciales.
	**SIN OBRAS DERIVADAS (*No Derivate Works*):** La autorización para explotar la obra no incluye la transformación para crear una obra derivada.
	**COMPARTIR IGUAL (*Share alike*):** La explotación autorizada incluye la creación de obras derivadas siempre que mantengan la misma licencia al ser divulgadas.

Mediante la combinación de estos cuatro preceptos se obtienen seis tipos de licencias:

- *Reconocimiento de autoría (CC BY)*
- *Reconocimiento de autoría – compartir en idénticas condiciones (CC BY-SA)*
- *Reconocimiento de autoría – Sin obra derivada (CC BY-ND)*
- *Reconocimiento de autoría – Sin uso comercial (CC BY-NC)*
- *Reconocimiento de autoría – Sin uso comercial – compartir en idénticas condiciones (CC BY-NC-SA)*
- *Reconocimiento de autoría – Sin uso comercial – Sin obra derivada (CC BY-NC-ND)*

Las licencias de la versión CC 4.0 abordan las características específicas de los datos.

Science Commons es una iniciativa dentro de Creative Commons que, entre otras cosas, pretender derribar barreras y desarrollar herramientas para facilitar la reutilización de datos resultados de proyectos de

investigación. En esta línea, Science Commons Open Access Data Protocol[32] recoge una metodología y unas buenas prácticas para la creación de herramientas que permitan la integración de bases de datos científicos entre sí y su puesta en el dominio público.

Siguiendo el modelo de Creative Commons, la Open Knowledge Foundation ha creado unas licencias específicas para colecciones de datos: *"The Open Data Commons Licence"*[33]. Es importante distinguir entre la licencia de los datos incluidos en la base de datos y el régimen de licencia de la base de datos en sí. Entre las licencias de Open Data Commons destacan las Database Contents License, referida a los contenidos de la base de datos, y la más radical de todas, la Public Domain Database License, en que los titulares de los derechos se despojan de ellos para beneficio de todos.

5.6 Preservación

Los datos deberán ser preservados y permanecer accesibles y utilizables para la investigación futura. La gestión de los datos deberá incluir un plan de preservación conforme a estándares internacionales.

Las cuestiones a plantearse son: ¿Qué datos hay que guardar? ¿Cómo guardarlos?

Hacer copias de seguridad de forma regular que puedan ser utilizadas para restaurar los ficheros originales. Es necesario constatar la integridad de los ficheros mediante la comprobación del código MD5 checksum value, el tamaño del fichero y la fecha.

La estrategia de almacenamiento de datos debe contemplar la obsolescencia del hardware y del software. Conviene copiar los datos en diferentes tipos de soporte físico, por ejemplo en uno digital y en un disco duro. Hay que tener en cuenta factores de conservación de los soportes, como los cambios de temperatura, la humedad relativa, la luz, etc.

[32] http://sciencecommons.org/projects/publishing/open-access-data-protocol/ [Consulta 12/12/2012]
[33] http://opendatacommons.org/licenses/ [Consulta 12/12/2012]

6. EJEMPLOS DE BUENAS PRÁCTICAS POR DISCIPLINAS Y ACTORES

6.1 Guías para la gestión de los datos:

- Australian National Data Service: HTTP://ANDS.ORG.AU/RESEARCHERS/MANAGE-DATA.HTML [Consulta 8/12/2012]
- Australian National University. Data Management: Information from courses and a manual on data management: HTTP://ILP.ANU.EDU.AU/DM/ [Consulta 8/12/2012]
- CIESIN: Geospatial Electronic Records- Resources on managing and preserving geospatial data and related electronic records: HTTP://WWW.CIESIN.COLUMBIA.EDU/GER [Consulta 8/12/2012]
- Data Management for Researchers: HTTP://ANDS.ORG.AU/RESEARCHERS/MANAGE-DATA.HTML [Consulta 8/12/2012]
- Gestión de datos en Humanidades: HTTP://ERCIM-NEWS.ERCIM.EU/EN89/SPECIAL/DATA-MANAGEMENT-IN-THE-HUMANITIES [Consulta 8/12/2012]
- ICPSR Guide to Social Science Data Preparation and Archiving: Outlines best practices throughout the research process, including applying for a research grant, collecting data, and preparing data for deposit in a public archive. HTTP://WWW.ICPSR.UMICH.EDU/FILES/ICPSR/ACCESS/DATAPREP.PDF [Consulta 8/12/2012]
- Oak Ridge National Laboratory. Best Practices for Preparing Environmental Data Sets to Share and Archive. Describes the practices to make data sets ready to share with others: HTTP://DAAC.ORNL.GOV/PI/BESTPRACTICES-2010.PDF [Consulta 8/12/2012]
- UK Data Archive: Create & Manage Data: Provides best practice strategies and methods for creating, preparing and storing shareable datasets. HTTP://WWW.DATA-ARCHIVE.AC.UK/CREATE-MANAGE [Consulta 8/12/2012]
- UK Data Archive: Managing and Sharing Data: a Best Practice Guide for Researchers 3rd. ed. HTTP://WWW.DATA-ARCHIVE.AC.UK/MEDIA/2894/MANAGINGSHARING.PDF [Consulta 8/12/2012

6.2 Datos por disciplinas:

- Anotación y Descripción de las bases de Datos Biomédicas (Harvard University): HTTP://ESCHOLARSHIP.UMASSMED.EDU/CGI/VIEWCONTENT.CGI?ARTICLE=1000&CONTEXT=JESLIB [Consulta 8/12/2012]
- Arqueología: HTTP://ARCHAEOLOGYDATASERVICE.AC.UK/ [Consulta 8/12/2012]
- Astronomía: HTTP://ADSWWW.HARVARD.EDU/ [Consulta 8/12/2012]
- Bioinformática: HTTP://WWW.EBI.AC.UK/INFORMATION/DATABASES_SITEMAP.HTML [Consulta 8/12/2012]
- Ciencias Marinas: HTTP://WWW.MARINE-GEO.ORG/CONTRIBUTE.PHP [Consulta 8/12/2012]
- Ciencias Químicas: HTTP://WWW.CHEMSPIDER.COM/ [Consulta 8/12/2012]

- Datos geoespaciales: HTTP://EDINA.AC.UK/PROJECTS/SHAREGEO/ [Consulta 8/12/2012]
- Energía: HTTP://EN.OPENEI.ORG/WIKI/MAIN_PAGE [Consulta 8/12/2012]
- ISATOOLS Datos de Biomedicina: HTTP://ISATAB.SOURCEFORGE.NET/ [Consulta 8/12/2012]
- Lingüística: HTTP://WWW.LANGUAGE-ARCHIVES.ORG [Consulta 8/12/2012]
- Listado de repositorios de Datos: HTTP://DATACITE.ORG/REPOLIST [Consulta 8/12/2012]
- Medicina: HTTP://WWW.NCBI.NLM.NIH.GOV/GENBANK/ [Consulta 8/12/2012]
- Música Brainz, HTTP://MUSICBRAINZ.ORG/ [Consulta 8/12/2012]

7. CASOS DE ESTUDIO EN ESPAÑA

En esta sección del informe se han reflejado todas las iniciativas de las que se tiene conocimiento sobre gestión de datos científicos en las que estén involucrados agentes españoles. La metodología seguida para elaborarlo es la siguiente: revisión bibliográfica, seguimiento de jornadas y conferencias, contactos con los grupos conocidos, revisión de proyectos y, por último, identificación de repositorios y datasets españoles en los bancos de datos registrados en el inventario internacional ODiSEA. Además se describe con detenimiento este proyecto como caso de estudio en nuestro país.

En primer lugar se detectaron **autores y temas de interés** a través de la literatura académica y profesional, de las que se da cuenta con detalle en el siguiente apartado. El primer autor que publica un trabajo sobre gestión de datos científicos en una revista del área de la información es Martínez Uribe en 2008, quien estuvo trabajando en Gran Bretaña. Durante los siguientes años se comienza a publicar y comunicar en foros públicos sobre la gestión de datos científicos y el compartido: el grupo de Granada, Torres Salinas, Robinson-García y Cabezas-Clavijo, (*Anuario ThinkEPI y El profesional de la información*), Pérez González desde Galicia (*Jornada SEDIC* y *75th Annual Meeting of the Society of American Archivists*) y en *Blok de BiD: reseñas de biblioteconomía y documentación* Melero y Peset de Valencia y Keefer y Borrego de Barcelona. En otras áreas, como la psicología (Botella Ausina y Ortego Maté) o las ciencias de la tierra (Bermúdez, Barragán y Alonso) también se localizan aportaciones particulares.

En segundo lugar, aparecen tres **reuniones** iniciales, cuyas aportaciones son detalladas en el apartado siguiente. La primera, promovida por FECYT-RECOLECTA y auspiciada por UNED en noviembre de 2011 (*Almacenamiento, la conservación y la gestión de los datos de investigación*), está orientada a la función de los repositorios con respecto a los datos. Como agentes activos españoles sólo aparece FECYT, y la biblioteca de datos del Centro de Estudios Avanzados en Ciencias Sociales, de la Fundación Juan March -presentada por Fernández y Martínez Uribe-, así como López Medina como coordinadora. Las otras tres fueron organizadas por GrandIR, *spin off* liderada por de Castro: la primera de agosto 2011 (*STM research data Management*), la segunda de mayo de 2012 (*Advances in research data management in Spain*) y la tercera de noviembre 2012 (*EuroCRIS autumn membership meeting*). En *STM research data Management,* investigadores de diferentes disciplinas exponen las formas en que gestionan los datos y sus necesidades; orientación que se afianza en *Advances in research data management in Spain*, junto con las perspectivas del Grupo de trabajo de Repositorios de datos y de los gestores de información de GrandIR, UPC y UOC.

Existen otras contribuciones dispersas en *Primeres Jornades sobre Gestió de la Informació Científica-JGIC, de abril 2012,* y el *5º Os-Repositorios* (mayo 2012) donde se identifican varias contribuciones y personas relacionadas con la gestión de datos. En octubre de 2012 se celebró *EUDAT European Data Infrastructure 1st Conference,* sin especial participación española (RedIRIS y Barcelona Supercomputing Center como socios).

En cuanto a **proyectos relacionados** y **contactos personales** con profesionales del sector de la gestión de la información se detectan dos claros grupos de interés en Barcelona y en Madrid. De los numerosos proyectos relacionados con datos en general, sólo algunos caerían dentro de la definición de "datos de investigación" utilizada en este informe: Wf4Ever, agINFRA y SeaDataNet. El resto, pareciendo muy cercanos, toman fuentes de datos no consideradas como "datos de investigación", lo que se ha comprobado contactando con sus directores (Baeza-Yates o Larriba). Por último, también detectamos algunas recomendaciones desde los Centros nacionales de referencia: Instituto de salud Carlos III (registro de biobancos) o el Centro Nacional de Datos Polares y Archivo Polar (Informe SCOR). Todo ello se revisa extensamente en el apartado "Evolución de las contribuciones españolas".

En último lugar, como se ha mencionado, se realizaron búsquedas en los **bancos de datos registrados en ODiSEA**. Se ejecutaron en campos generales, de autor o geográficos, si los tenían, por los términos Spain o Spanish. De este trabajo puede concluirse lo siguiente: i) de los 183 items de ODiSEA revisados (1-15 octubre) sólo se han identificado dos registros españoles: CEACS -la biblioteca de datos de la Fundación Juan March- y Digital.CSIC –repositorio institucional del CSIC, desarrollado en Dspace-; ii) la mayoría de los bancos no permiten restringir por país; iii) las búsquedas ofrecen resultados ambiguos y escasos: aproximadamente sólo 20 ofrecían conjuntos de datos españoles. Si bien la cifra es verdaderamente baja, quizá el número de *datasets* depositados pueda ser mayor. No se ha realizado este análisis ya que la única intención en estos momentos ha sido comprobar que existe actividad con respecto a datos por parte de los investigadores españoles; iv) se observa que las únicas materias representadas son Mineralogía, Sociales, Ciencias de la tierra y de forma preponderante Biomedicina en sus varias ramas. Otros expertos mencionan también la presencia española en Historia Económica, Química y Climatología.

7.1 Evolución de las contribuciones españolas. Gestión de datos científicos

7.1.1 Revisión bibliográfica de literatura académica y profesional

- Martínez Uribe y Macdonald (2008) comentan la revolución que suponen los movimientos abiertos en la comunicación científica, las nuevas formas de trabajo científico (e-ciencia), y los nuevos roles que han de asumir las unidades de información como entidades de preservación. Acuñan el término ***data curation*** como la gestión de los datos durante todo su ciclo de vida, desde su creación hasta la adición de valor para su reutilización. Citan las iniciativas más relevantes, especialmente las británicas: tanto líneas de estudio como resultados concretos.
- Torres Salinas (2009) aborda los **beneficios y formas de compartir datos**, así como la **iniciativa DAF** en sendas Notas ThinkEPI. En la primera revisa bibliográficamente las opiniones de los propios científicos y de las agencias financiadoras gubernamentales –mandato del NIH-, para acabar señalando que se trata de una nueva oportunidad para las bibliotecas académicas, al igual que lo fue el movimiento de acceso abierto a las publicaciones. La segunda de sus notas describe el

Data Audit Framework (DAF) como hoja de ruta posible para cualquier institución que se plantee gestionar datos. El objetivo de DAF es conocer la situación para plantear mejoras. Los resultados de las auditorías en Edinburgh, Bath, Glasgow, King´s College, Southampton y Oxford revelan una gestión "casera" que necesita de políticas institucionales para preservar los datos, así como guías para ayudar a los propios investigadores.

- Melero (2010) reseña *Riding the Wave*, informe clave encargado por la Comisión Europea a un grupo de expertos para conocer los **beneficios y coste** de la puesta en marcha de una **infraestructura global de datos fiable y estable**. En él se considera vital: no perder la flexibilidad; crear incentivos para compartir sin perder la privacidad; preservar los datos enriqueciéndolos con su contexto y procedencia; y los modelos de financiación. Proponen entre otras acciones, con vistas al 2030: crear una infraestructura colaborativa internacional de datos con fondos adicionales; medir y recompensar el valor de estos datos; o crear un comité internacional e interministerial para dirigir esta infraestructura. Identifica algunos sectores muy necesitados de esta globalización de datos: cambio climático y medio ambiente, temas energéticos, epidemiológicos, etc. Sitúa la principal preocupación en cómo integrar las distintas partes del proceso e incentivar la participación.
- Keefer (2011) analiza el informe de Itaca S&R sobre el papel de las agencias de financiación en la sostenibilidad de los datos. Basado en una encuesta a 25 europeas y norteamericanas, detecta una falta de uniformidad en sus procedimientos. No obstante, destaca el papel que pueden asumir estas agencias si exigieran un plan de gestión de datos que asegurara su preservación.
- En 2012 Torres Salinas, Robinson y Cabezas reúnen los aspectos más reseñables del ***data sharing*** con un enfoque híbrido, vinculando nuestra profesión con las tendencias que se producen al mundo de la investigación, que al fin y al cabo es donde el compartido de datos comienza. Ofrece una detallada discusión del propio concepto datos de investigación, de las formas de compartir y los bancos disponibles, así como de las políticas de las agencias financiadoras y de las revistas.

  Borrego (2012a) reseña un informe procedente del proyecto ODE-Opportunities Data Exchange sobre la **integración de datos primarios y publicaciones** de forma que no se pierda su relación. El informe recoge ejemplos y pone de manifiesto el deseo de reutilizar datos ajenos pero cierta reticencia a compartir los propios aduciendo problemas legales. Las vías preferidas de almacenamiento son los repositorios y las plataformas editoriales, aunque la realidad muestre poca actividad en ambos casos. Relacionarlos con las publicaciones cuenta con ventajas adicionales: ayuda a interpretar los datos y proporciona valor tanto a los investigadores que los comparten como a las propias publicaciones. Validación y preservación son los problemas que detectan si quedan en manos de las editoriales. Por último, señala cómo en el proceso investigador, los centros de cálculo -encargados de la recogida y el procesamiento de los datos- y las bibliotecas –donde se almacenaban las publicaciones- parece que van a adoptar papeles complementarios gracias a la gestión de datos.

- Peset (2012) reseña los resultados de una encuesta sobre la información científica en la era digital. Frente al acceso a las publicaciones, el acceso a los datos científicos es percibido como más problemático debido a la falta de infraestructuras, de incentivos y de políticas nacionales. Cabe resaltar que para los productores de datos el mayor problema son los incentivos, mientras que para los gestores es la infraestructura.
- Borrego (2012), por último, da cuenta de los resultados de cuatro informes: los tres primeros de ámbito europeo (procedentes de las iniciativas KE y ODE) y un cuarto de alcance norteamericano, procedente de Council of Library and Information Resources-CLIR.
    - El primero, para estudiar una posible infraestructura a escala europea, ya ha sido mencionado por Castro. Se centra en el análisis de los incentivos a los creadores de datos, de las iniciativas de formación y por último en las características y necesidades de financiación de la infraestructura tecnológica a partir de cuatro ejemplos.
    - El segundo informe, basado en una encuesta a bibliotecarios, aborda las implicaciones de una correcta citación de los datasets, que entre otras cuestiones, permitiría cuantificar su uso. En todo caso, pone de manifiesto la escasa demanda a las bibliotecas de servicios sobre estos datos publicados como material adicional a los artículos de los investigadores.
    - El tercer informe, también derivado del trabajo de ODE se basa en entrevistas a expertos, que identifican las cuestiones clave: rol de los editores, modelos de financiación, necesidad de formación y estándares específicos, etc.
    - Finalmente, el cuarto procede de CLIR y está fundamentado en un estudio cualitativo sobre investigadores del área de ciencias sociales y en el análisis de la oferta formativa en gestión de datos. Entre los científicos destaca el escaso interés que suscita por qué no se incentiva y la dificultad de gestionar los datos debido a la complejidad de su ciclo de vida. Mientras que la última parte del trabajo identifica muy pocos centros que ofrezcan esta formación y en todo caso, siempre en un nivel avanzado de estudios.
- Desde otro ámbito, la psicología, Botella Ausina y Ortego Maté (2010) proponen **vías de actuación** ante la especial reticencia a compartir datos que detectan en los científicos de su sector. Caracterizan los métodos y la naturaleza epistemológica de la psicología, lo que incide en la escasa costumbre de compartir. Resumen los beneficios en su campo porque el compartido "es un principio ético, ayuda a prevenir el fraude y protege contra algunas amenazas a la fiabilidad" (p. 266). Recomiendan crear las herramientas que permitan compartir y establecer alianzas estratégicas con personas e instituciones líderes de su disciplina.
- Desde el sector de la investigación antártica, Bermúdez, Barragán y Alonso (2011) presentan la **política** y el **protocolo PRADDA**, http://hielo.igme.es del Centro Nacional de Datos Polares, desde el que se comparte información con los firmantes del Tratado Antártico. Este sector muestra ser pionero a causa de su internacionalización y la creación en 1989 del *Committee in the Coordination of the Antarctic Data-CCAD*. Desde 2005 los proyectos financiados con el Plan Nacional I+D+i están

obligados a enviar copia de sus datos al Centro Nacional de Datos Polares, que quedan inexorablemente a libre disposición en 4 años aproximadamente.

7.1.2 Jornadas y conferencias relacionadas con la gestión de datos de investigación

- Durante los **webinars** de RECOLECTA (2011), FECYT se presenta como socio español del proyecto OpenAire plus (hasta 1 junio 2014) cuyos objetivos son incluir datos, vincularlos a las publicaciones y generar servicios adicionales para la comunidad investigadora.
  Por otro lado, Fernández y Martínez Uribe dan a conocer un caso interesante de estudio, la colección digital de datos y libros de códigos recogida en la biblioteca del CEACS. Desde 1987 recopila microdatos, datos agregados y geográficos, ingresados por compra, por alianzas estratégicas (ICPSR) o producidos por los investigadores sobre opinión, sistemas políticos, elecciones, encuestas sociales, datos geográficos y españoles. Destacan tres núcleos de servicios especializados al investigador:
  - Colabora en la selección y compra de los datos en diferentes formatos, los describe y ofrece un servicio de referencia, así como participación en el establecimiento de licencias de acceso o en la comunicación con otros centros;
  - Ayuda al investigador en el uso de las aplicaciones de análisis estadístico, de visualización, y de utilidades para extraer datos (data scraping);
  - Asesora a los investigadores desde el momento de la creación de datos hasta su depósito para su preservación (selección de formatos, elaboración de libros de códigos, almacenamiento en la Fundación y en Dataverse…).

  La utilización de esta última aplicación de código abierto de la Universidad de Harvard tiene varios beneficios, entre los que destacan los siguientes: trabaja con identificadores persistentes y esquemas normalizados de metadatos para ciencias sociales (DDI) y genera automáticamente las citas en formatos estandarizados. Destaca la relación estrecha con los investigadores y su contribución a crear entre ellos y los estudiantes una cultura en el manejo de datos.
- Estrada y Echenique (2011) muestran durante la **jornada STM** el estado de las bases de datos en química cuántica y la necesidad del proyecto Quixote ante la falta de un modelo de datos estandarizado en la disciplina del cálculo. En 2010 proponen la gestión de los datos derivados de los cálculos del tamaño de las moléculas pequeñas y medianas. Desarrollan una infraestructura tecnológica para estandarizar y enriquecer semánticamente los formatos de los datos, que puede ser implementada a nivel individual, de proyecto, etc. En esta primera jornada los participantes concluyen la necesidad de:
  - Constituir un grupo de trabajo interdisciplinar.
  - Analizar cómo los investigadores STM almacenan sus datos internacionalmente.
  - Concienciar para compartir los datos.

- Promocionar las iniciativas de eSpacio UNED y Digital CSIC.
- Estudiar si se están depositando en las plataformas de las editoriales.
- Crear un protocolo en los grupos de investigación.
- Incentivar el compartido de datos (por ejemplo reconociéndolo como contribución científica
- Por último, detectan grandes diferencias entre disciplinas y ponen de manifiesto el potencial de las bibliotecas y la Red Española de e-Ciencia del MICINN como soporte a estas actividades.

- Pérez González (2010) durante la XII Jornada SEDIC presenta **tres modelos de coste** del mundo anglosajón para planificar una política de preservación de la información digital en general, y los datos en particular, dada la obligatoriedad de depositarlos en Gran Bretaña. El primero deriva del *Blue Ribbon Task Force on Sustainable Digital Preservation and Access,* fundación participada por organismos públicos y privados de EEUU y Gran Bretaña. Su enfoque es general, aunque entre la información digital a preservar menciona el discurso académico y los datos de investigación. Pone de manifiesto que el problema ha de ser abordado desde una perspectiva de coordinación entre múltiples agentes. El segundo modelo, *Keeping Research Data Safe*, proviene del JISC. Estudia tres casos, Cambridge, King's College y York University, para establecer sus recomendaciones basadas en OAIS. Destaca el detalle exhaustivo de los beneficios en varios niveles. El último de los modelos, LIFE, se basa en el ciclo de vida propuesto por la *British Library*, con OAIS de nuevo. En función del tamaño y propósito del archivo implementa en plantillas Excel unos modelos para obtener un resumen de costes para la institución.
- Pérez González (2011) en el 75th Annual Meeting of the Society of American Archivists presenta las **bases de un proyecto de gestión** de datos a nivel autonómico en Galicia, **basado en buenas prácticas internacionales** entre las que cita expresamente a Holanda, Estados Unidos y Gran Bretaña. Su planteamiento tiene como base el ciclo de vida de los datos, desde la creación a la conservación y los servicios.
- En JGIC, Peset menciona el trabajo de **datos de investigación** distinguiendo los datos producidos por los investigadores -pequeños conjuntos de datos que pueden almacenarse en las plataformas editoriales- y los datos de los grandes productores generalmente gubernamentales, que en ocasiones son semánticamente compatibles.

  El resto de contribuciones que abordan datos en esta jornada son enfoques orientados hacia la **medición de los sistemas científicos y la evaluación de la ciencia**, algo que se aleja del objeto de estudio del presente informe: descripción de los sistemas estadísticos que recogen datos sobre la ciencia (UNEIX y Institut català d'estadística) o el análisis bibliométrico de datos (Borrego).
- La segunda jornada **Research Data Management** (mayo 2012) reúne investigadores y gestores de información.

  Sorribas muestra el trabajo de la Unidad de Tecnología Marina (UTM) del CSIC como apoyo a la investigación marina y polar. Trabajan en varios proyectos de gestión de datos a escala nacional e internacional: RedICTS, EuroFLeets, ICOS, ESONET, CMIMA y OGC. En 2008 el grupo de trabajo

SCOR España (Comité Científico sobre Investigación Oceánica) realiza un análisis DAFO en "*Reflexiones sobre la gestión y custodia de datos oceanográficos en España. Recursos existentes y recomendaciones para el futuro*". Destaca sus obligaciones con el entorno internacional y que los datos marinos son complejos por su escala, temática, instrumentos..., pero cuentan con un modelo de datos, con vocabularios especializados, etc.

Por su parte, Lahoz, fonetista en la UPM detecta la ausencia de un repositorio de datos en bruto que admita entre sus metadatos la representación de alguna singularidad de su sector.

Vallverdú presenta el caso del departamento de Teoría de la señal y comunicaciones (UPC) en el que ya están almacenando y procesando datos de señales.

La representación de los gestores de información en esta ocasión es bastante nutrida.

Castro muestra la reciente sensibilidad sobre el manejo de datos y varios proyectos nacionales integrados en la iniciativa Knowledge Exchange-KE: británico (JISC-MRD), alemán (DFG), holandés (SURF) y danés (DK). Remarca la necesidad de:

- Reconocer el compartido de datos mediante incentivos
- Facilitar programas de formación después de identificar los grupos de interés entre editoriales, investigadores, evaluadores de proyectos, gestores de información…
- Estudiar la infraestructura necesaria, tomando en cuenta los casos previos de éxito y distinguiendo los retos derivados del trabajo con datos de los derivados de la implementación técnica.
- Encontrar modelos de coste adecuados para tres grandes áreas: ciencias físicas, ciencias biomédicas y ciencias sociales y humanidades. Muestra algunos casos de éxito de Dspace: el repositorio institucional Edinburgh DataShare, LAGO en Colombia y Dryad.

Concluye, entre otros aspectos, que "No es imprescindible contar con un repositorio de datos para comenzar a planificar estrategias de tratamiento de los datos".

Por su parte, Serrano presenta los avances del Grupo de trabajo FECYT/RECOLECTA de repositorios de datos.

Zúñiga muestra las iniciativas de la UOC en la gestión de datos, vinculadas al CRIS y el repositorio, pero muy pendiente de la perspectiva del investigador.

Por último, FECYT detalla OpenAirePlus, con los siguientes objetivos: enlazar publicaciones, sets de datos y fuentes de financiación para obtener "publicaciones mejoradas" (Enhanced Publications) implementando OAI-ORE. Aspira a ofrecer más información de contexto de muy diferentes clases (desde copyright a la comunidad a la que se dirige) para dos disciplinas piloto: ciencias sociales y humanidades y ciencias de la vida.

- En **OS-Repositorios** Castro, García y Rodríguez Miranda presentan una visión panorámica internacional de los avances en gestión de datos y el proyecto ADDI Laboratorio de documentación geográfica del patrimonio, UPV/EHU); y García García y Rodríguez Gairín un avance de resultados de ODiSEA, que más abajo describimos extensamente.

7.1.3 Proyectos relacionados con la gestión de datos y contacto con profesionales del sector

En cuanto al tercer método para identificar otros agentes, ha sido la búsqueda en nuestro país de proyectos relacionados y los contactos personales con profesionales del sector de la gestión de la información. Se detectan los siguientes **grupos de interés** en Barcelona y Madrid.

- i-VIU: Grup d'estudis mètrics sobre el valor i ús d'informació en la Universitat de Barcelona (Borrego y Urbano), trabaja en estudios métricos y estadísticos de información en el entorno digital.
- En el CESCA (Ricard de la Vega) trabajan en infraestructuras y plataformas de búsqueda y difusión de la actividad científica -por ejemplo datos transmitidos entre el CERN y la Anella Científica-. La Anella Científica es una red de comunicaciones de alta velocidad, que permite a los científicos volver a acceder a los datos acumulados de los experimentos para su análisis.
- En la Universidad Carlos III de Madrid el grupo Tecnodoc (Méndez, Gómez y Hernández) abarca actividades relacionadas con la comunicación científica y la medición de la investigación y contaron el año pasado en la cátedra de excelencia con Jane Greenberg (Dryad).

Por otra parte, existen múltiples **proyectos relacionados con datos** en general. Algunos de ellos han de tomarse en consideración en este informe ya que trabajan sobre datos de investigación.

- De la Universidad Politécnica de Madrid-UPM, Instituto de Astrofísica de Andalucía e ISOCO, encontramos el proyecto Wf4Ever, cuyo trabajo versa sobre la evolución, intercambio y colaboración de flujos de trabajo. El objetivo principal es proporcionar los medios adecuados para maximizar la participación y la reutilización de los Objetos de Investigación conservados, mientras se da soporte a su evolución y versionado y se facilita la colaboración entre los científicos.
- Otro proyecto internacional, agINFRA "Promoting data sharing and development of trust in agricultural sciences", cuenta con la Universidad de Alcalá de Henares como miembro español. Consiste en implementar una infraestructura abierta para mejorar la transferencia de resultados científicos y tecnológicos del área de la agricultura estableciendo además normas de intercambio, programas y metodologías.
- La Red Paneuropea de Gestión de Datos Marinos y Oceánicos (SeaDataNet) está participada por el Centro Español de Datos Oceanográficos del Instituto Español de Oceanografía. Desarrollan un sistema interoperable para la gestión de datos e información marina que se han comentado anteriormente. Además, plantean que es necesario elaborar una normativa que regularice los derechos y las obligaciones de los generadores de los datos e información así como las condiciones de uso de dicha información.
- En el Instituto de Física de Cantabria (IFCA) del CSIC (Matorras) los grupos de investigación de Computación Avanzada y e-Ciencia[34] y el de Física de Partículas participan en el proyecto del

[34] http://www.ifca.unican.es/computacion_avanzada_y_e-ciencia [Consulta 15/12/2012]

CERN Large Hadron Collider. En lo que se refiere a datos puros, el Worldwide LHC Computing Grid[35] forma parte de esta iniciativa de Física de Partículas. Se trata de un proyecto internacional que conecta más de 170 centros de computación GRID en 36 países que almacena, distribuye y analiza unos 25 millones de Gygabytes generados cada año por el gran colisionador,

- Desde el Instituto Pirenaico de Ecología (IPE) y la Escuela Experimental de Aula Dei (EEAD) del CSIC, Vicente-Serrano y Begueria han desarrollado una base de datos **SPEIbase** que recoge mensualmente datasets de sequía a escala global basados en un nuevo índice de sequía, el Standardised Precipitation-Evapotranspiration Index para el período 1901-2006. Se puede consultar desde una interfaz de búsqueda propia, aunque el almacenamiento y gestión de los datos se realiza en el repositorio Digital.CSIC.
- También un grupo de investigadores del Instituto de Historia (IH) del Centro de Ciencias Humanas y Sociales (CCHS) del CSIC compuesto de Crespo, Pérez, Maestre, y del Bosque desarrollan DynCoopNet -Dynamic Complexity of Cooperation-Based Self-Organizing Commercial Networks in the First Global Age-, base de datos que recoge datasets sobre mercaderes, agentes financieros, compañías de monopolio, rutas comerciales, etc. de 1400 a 1800, de Europa atlántica, mundoatlántico-americano y Asia-Pacífico. Disponible desde Digital.CSIC.
- Otros proyectos, pareciendo muy cercanos, toman fuentes de datos no consideradas como "datos de investigación", lo que se ha comprobado examinando sus web o contactando con sus directores: en la Universitat Pompeu Fabra (Baeza-Yates), la Web Research Group está desarrollando el proyecto Plataforma Modular y Extensible para Minería de Datos en la Web**,** para todo tipo de datos disponibles en la web. Entre los objetivos planteados para esta infraestructura se encuentran: la recolección y extracción de datos de la Web, el almacenamiento de objetos (datos, metadatos, vistas y relaciones). Los procesos que incluyen operaciones de recuperación de información, de similitud, de relevancia total o parcial (ranking), de manipulación de objetos, grafos y secuencias temporales, y de estadística, que permiten generar nuevas vistas sobre los datos y relaciones lógicas entre ellos, y la presentación de vistas y relaciones usando técnicas de visualización para datos estructurados.
- En la Universitat Politècnica de Catalunya (Larriba), el DAta MAnagement group (DAMA-UPC) grupo de investigación para la gestión y análisis de grandes grupos de datos, ha desarrollado **Science-a** [36] [Consulta 9/12/2012] , una solución para el tratamiento y visualización de los datos, y el soporte a la elaboración de proyectos. Toman los datos de Cordis, el Servicio de Información Comunitario sobre Investigación y Desarrollo, que informa de todos los proyectos europeos que ha habido hasta el momento. Gestiona toda la información de los proyectos, con abstracts, Partners, cantidad de dinero asignada, etc.

  Otros muchos proyectos examinados han sido directamente descartados para este informe, muchos de ellos de carácter tecnológico.

[35] http://wlcg.web.cern.ch/ [Consulta 15/12/2012]
[36] http://www.sciencea.com [Consulta 15/12/2012]

- **ADMIRE-*Advanced Data Mining and Integration Research for Europe*** (UPM) está motivado por la dificultad de extraer información significativa realizando minería de datos, combinaciones de múltiples recursos heterogéneos y distribuidos. ADMIRE propone una arquitectura en lenguaje DISPEL para expresar los flujos de trabajo de minería de datos y la integración a través de los perfiles de usuario.
- El proyecto **PlanetData**, con participación de la UPM, establece una comunidad europea sostenible de investigadores que apoyen a las organizaciones en la publicación de sus datos en nuevas y útiles formas, aumentando la habilidad de las organizaciones de dar sentido a las enormes cantidades de datos publicados online de forma continua. Incluye datos estructurados y no estructurados, flujos de datos, (micro) entradas del blog, archivos digitales, recursos de e-Ciencia, conjuntos de datos del sector público, y datos enlazados de "la nube".
- Otro proyecto de la UPM, de carácter nacional es **MyBigData**, cuyo objetivo es crear una plataforma que integre nuevos métodos, técnicas y herramientas para permitir la integración de fuentes de datos heterogéneas de carácter científico mediante el uso de ontologías, incorporando nuevos tipos de fuentes de datos procedentes de redes sociales de investigadores, de la Web de Linked Data, y de redes de sensores.
- El **proyecto BabelData** (UPM) tiene como objetivo desarrollar técnicas y algoritmos para la construcción de servicios que sean capaces de crear y utilizar ontologías y datos multilingües. Para lograr este objetivo, el proyecto plantea abordar los siguientes aspectos:
  - Localización automática de ontologías
  - Mappings multilingües entre ontologías en distintas lenguas
  - Modelos, métodos, técnicas y herramientas para la generación de datos enlazados multilingüe
  - Servicios para la integración de ontologías multilingües y Linked Data multilingüe
- El **proyecto GeoBuddies** (UPM) ha desarrollado una aplicación que proporciona información y servicios geo-espaciales, para los peregrinos del Camino de Santiago. La aplicación soporta el acceso y la interacción móvil, dinámica y dependiente de contexto con un conjunto de recursos y servicios, y toma los datos del Instituto Geográfico Nacional (IGN).

Por último, también se detectan algunas recomendaciones desde los centros nacionales de referencia como el Instituto de Salud Carlos III. Con el apoyo del Ministerio de Economía y Competitividad, pone a disposición de los investigadores una plataforma electrónica para el registro de biobancos y colecciones de muestras, para facilitar la consulta pública y el acceso a los materiales que albergan.

El Centro Nacional de Datos Polares Español (CNDP), en el Instituto Geológico y Minero de España (IGME), se ocupa de la generación de metadatos y del almacenamiento, custodia y gestión de los datos brutos derivados de la los proyectos de investigación. En el marco Subprograma Nacional de Investigación Polar (SNIP) el CNDP elaboró una política de datos en 2007: "Propuesta de protocolo de remisión, almacenamiento y difusión de datos antárticos en España". PRADDA define su ámbito de aplicación, los

tipos y formatos de datos, los procedimientos de envío de datos al CNDP y la accesibilidad y disponibilidad de los datos mediante solicitud de acceso controlada.

Otras recomendaciones vienen dadas por comités nacionales como el Comité Científico sobre Investigación Oceánica SCOR-España, en su informe "Reflexiones sobre la gestión y la custodia de datos oceanográficos en España, Recursos existentes y recomendaciones para el futuro". Para corregir algunos de los problemas de la gestión de datos oceanográficos en España, recomienda que el sistema de gestión de datos oceanográficos debiera garantizar los siguientes servicios:

- La recopilación, el control de calidad y el almacenamiento de datos de manera que queden disponibles para el futuro. Para ello es necesario establecer:
  - ✓ Rescate de datos y metadatos que no estén accesibles actualmente.
  - ✓ Política de datos que contemple:
    - Obligatoriedad de depositar datos generados con dinero público en "Centros Acreditados".
    - Fuentes de financiación para asegurar la gestión de datos.
    - Normativa sobre restricciones y permisos de uso.
  - ✓ Una estructura que permita la coordinación e integración de la información de los centros que funcionan actualmente y que ofrezca servicios de datos a los miembros de la comunidad científica que lo requieran.
- Distribución de datos a científicos, gestores, industria y público: facilitar el acceso a los datos de las estaciones meteorológicas y oceanográficas operativas en aguas españolas. Para ello es necesario coordinar e integrar las bases de datos existentes, de manera que los usuarios puedan localizar la información necesaria a través del Centro Coordinador.
- Establecimiento de protocolos para la adquisición y tratamiento de datos: es necesario seleccionar los protocolos adecuados y diseminar la información pertinente a todos los investigadores interesados.
- Desarrollo de productos de datos adecuados a la evolución de la demanda.
- Adopción de una política de datos común nacional de obligado cumplimiento. Por otra parte, la adquisición de datos representa un trabajo que tiene que ser reconocido.

8. CASO DE ESTUDIO: ODISEA[37]

8.1 Antecedentes

Como se ha detallado, el rol de las editoriales en el acceso a los datos de investigación asociados a las publicaciones es un aspecto clave en su difusión que otorga grandes ventajas según los trabajos de OpenAirePlus y Opportunities data Exchange (Fecyt 2012, Borrego 2012a).

Varios investigadores -Univ. de València, Univ. de Barcelona, Univ. Politècnica de València, Univ. Ramón Llull, Univ. de Murcia y Univ. Católica de Valencia- de diferentes perfiles (salud, actividad física, documentación y nanotecnología) han trabajado en la identificación de las pautas de los investigadores a la hora de depositar los datos derivados de la investigación a partir de la presentación en FECIES 2011 y 2012 (Peset y otros, Ferrer-Sapena y otros). Este trabajo inicial se basaba en el análisis de las plataformas de las editoriales que eran usadas por las revistas con mayor factor de impacto de todas las disciplinas. Unos resultados previos sobre el área de sociales y humanidades fueron presentados en 2012 en la 2ª Conferencia sobre calidad de revistas de ciencias sociales y humanidades CRECS (García-García y otros), lo que permitió conocer la terminología utilizada y clasificar las editoriales estudiadas en función del nivel de reutilización que permitían.

De los numerosos procesos que tienen lugar durante la gestión de todo el ciclo de vida de los datos científicos el almacenamiento en depósitos cobra una gran importancia por sus implicaciones en cuanto a la preservación para el futuro, la citación, el reconocimiento de la autoría, la utilización de enlaces persistentes... Por esta razón, se indagó sobre la existencia de un registro mundial de depósitos de datos de investigación, tal y como existe desde hace años para repositorios de acceso abierto: ROAR y OpenDoar. No se encontró nada similar a nivel nacional ni internacional excepto recientemente Databib, con quienes se ha contactado para conocer su metodología y coordinar los proyectos.

La ausencia de un sistema central que los recopile ha conducido a que surja la necesidad de establecer un registro que los recoja y clasifique. La urgencia de esta acción responde a la proliferación de depósitos específicos en distintas disciplinas y a la descentralización de los depósitos de almacenamiento de datos en los repositorios de las propias instituciones. En estos momentos iniciales en la gestión de datos de la investigación se necesitan instrumentos básicos como ODiSEA que permitan ofrecer un panorama global de la situación que coadyuve a responder a los numerosos interrogantes que se identifican.

8.2 Objetivo

El grupo de investigación planteó la creación de una herramienta que aglutinara estos depósitos de forma clasificada por disciplinas. El objetivo de este proyecto es facilitar la identificación de las fuentes de

[37] http://odisea.ciepi.org/ [Consulta 15/12/2012]

almacenamiento de datos de investigación para permitir, como mínimo, a los profesionales de la información conocer de forma fácil y fiable dónde los investigadores deben depositar sus datos y si existen lagunas disciplinares.

Está previsto que ODiSEA tenga también las siguientes utilidades para el usuario y para los investigadores: i) proporcionar datos básicos sobre los bancos registrados, especialmente su grado de apertura; ii) ser el germen del futuro meta-buscador OPENDATASCIENCE de conjuntos de datos cosechables; y iii) crear conocimiento de investigación sobre la gestión y producción de los datos de investigación.

8.3 Equipo

El equipo está conformado por nueve personas, investigadores de las siguientes universidades:

- Alicia García-García de la Universidad Católica de Valencia
- Antonia Ferrer-Sapena de la Universidad Politécnica de Valencia
- Fernanda Peset de la Universidad Politécnica de Valencia
- José Morales-Aznar de la Universitat Ramon Llull
- Josep-Manuel Rodríguez-Gairín de la Universitat de Barcelona
- Luís-Millán González de la Universidad de Valencia
- Tomás Saorín de la Universidad de Murcia
- Xavi García –Massó de la Universidad de Valencia
- Florencia Atalia Dieci de la Universidad Politécnica de Valencia (Colaboradora)

Actualmente está financiado por el Plan Nacional de I+D+i del Ministerio de Economía y Competitividad: "OPENDATASCIENCE, centro de recursos para la preservación y gestión de datos abiertos de investigación", CS02012-39632-C02-02, y constituirá una parte del sitio web público del proyecto.

8.4 Metodología

La metodología seguida para recopilar los depósitos y repositorios de datos que existen se ha basado en varias fuentes. Se han revisado estudios bibliográficos previos interrogando las bases de datos de la Web of Knowledge, Scopus, CSIC y LISA, combinando las palabras clave: "data sharing", "reuse", "data curation", "research data" y "data repositories". Estos trabajos citaban depósitos como DART (Treloar, 2006), ARROW (Payne y Treloar 2006), DRYAD (Greenberg, 2009), Protein Data Bank y GenBank (Martínez-Uribe y Macdonald, 2009) y otros mencionaban conjuntos de ellos (Torres-Salinas, 2012).

Se ha observado que en los últimos años las editoriales han promovido la recepción de estos datos y la mayoría de ellas contemplan en su política para el autor unas pautas para el material complementario. Se han examinado las políticas de copyright de las editoriales científicas más relevantes, con respecto al material suplementario de los artículos, ya que en áreas como la Medicina o las Ciencias Naturales las

editoriales especifican los repositorios públicos en los que se deben depositar los conjuntos de datos para que el artículo pueda publicarse.

También se ha realizado la consulta de los registros de repositorios de acceso abierto ROAR (*Registry of Open Access Repositories*) y OpenDoar (*Directory of Open Access Repositories*) e identificado los archivos digitales que contienen datos de investigación.

La clasificación de los bancos de datos está basada en las áreas de conocimiento del *Essential Science Indicators* de la *Web of Knowle*dge: Agricultural Science, Biology and Chemistry, Chemistry, Clinical Medicine, Computer Science, Economics and Bussines, Engineering, Environment Ecology, Geoscience, Immunology, Material Science, Mathematics, Microbiology, Molecular Biology and Genetics, Multidisciplinary, Neuroscience and Behaviour, Pharmacology.

La web ODiSEA se desarrolló sobre Drupal, dentro del dominio CIEPI, y la herramienta de registro y búsqueda la administra Rodríguez Gairín en un hosting propio.

8.5 El producto: "ODiSEA: International Registry on Research Data"

DiSEA International registry on Research Data

Actualmente cuenta con 183 depósitos, entre los se encuentran bancos especializados, bibliotecas de datos, repositorios que aceptan conjuntos de datos, y bancos de imágenes.

Este registro permite buscar entre los depósitos que están registrados por nombre, por áreas geográficas, por disciplinas científicas… Y se continúa investigando para proporcionar resultados a medida que el acceso a los datos va siendo depositado en abierto y sobre las políticas de reutilización.

Se recoge la siguiente información de los depósitos de datos, aunque no toda está accesible al público: Nombre, Institución, Tipos de datos que almacena, Formato, Disciplina, Área geográfica, URL, Año, Cantidad de datos, OAI-PMH URL, Acceso abierto, Observaciones/notas: en el que se incluye software y otros datos.

8.6 Lecciones aprendidas

Durante la elaboración de ODiSEA se ha percibido una serie de denominadores comunes de naturaleza cualitativa que reseñamos como problema en la gestión de datos:

- Cada uno de los depósitos está recogiendo los datos de formas muy diferentes, más vinculadas a cada uno de los sectores disciplinares que a un marco común de trabajo.
- La información que ofrece cada uno de los ítems en su web no es homogénea, no sigue un patrón estandarizado.

- En muchas ocasiones la información que ofrecen como resultado, los datasets, no son comprensibles más que para los especialistas del ramo.
- Los webs no ofrecen suficiente información para los usuarios ajenos al sistema.

- No es posible conocer fácilmente el número de datasets introducidos en cada item, tampoco los registrados en DOAR.

9. BUENAS PRÁCTICAS

Por último, se reseñan algunas buenas prácticas y ejemplos detectados en los ítems registrados:

- Cantidad de Partners involucrados en el proyecto: Dryad cuenta con 23 socios del más alto nivel entre sociedades científicas, revistas, editoriales…
- Grado de apertura: la American Association for the Advancement of Science, de su revista Science permite explícitamente a cualquier usuario descargar, imprimir, extraer, reutilizar, archivar y distribuir los datos asociados a los artículos
- Estandarización de los datos: según gráfico de Cyganiak [38] algún depósito también expone los datos en la web semántica mediante linked open data como UniProt
- Visibilidad de los bancos: ciertos items logran alcanzar "fama" entre los autores españoles, lo cual es altamente recomendable para su difusión: Dryad, Archer, DART, GenBank…
- Coordinación a escala nacional: agencias como el National Institutes of Health estadounidense ha liderado proyectos sectoriales para promover estándares para datos: NINS Common Data Elements[39]
- Recomendaciones de las grandes editoriales: algunas indican que se han de depositar en bancos públicos específicos en función de su tipología y disciplina. Por ejemplo, la American Association for the Advancement of Science recomienda el depósito de estructuras moleculares de datos en Worldwide Protein Data Bank, secuencias de proteínas y ADN en GenBank, EMBL o DDBJ y microarrays en Gene Expression Omnibus y ArrayExpress

[38] http://richard.cyganiak.de/2007/10/lod/ [Consulta 9/12/2012]
[39] http://www.commondataelements.ninds.nih.gov/General.aspx#tab=Data_Standards [Consulta 9/12/2012]

10. SOBRE LOS CASOS DE ESTUDIO EN ESPAÑA

Es oportuna una estrategia nacional en nuestro país, al nivel de autoridad más alto posible y con el mayor número de agentes involucrados. Será necesaria una estructura que permita la coordinación e integración de la información de los centros que funcionan actualmente y que ofrezca servicios de datos a los miembros de la comunidad científica.

Ha de promoverse esta problemática especialmente entre **investigadores**, quienes tienen una vinculación vertical con expertos de su propia disciplina, a escala nacional e internacional. Es imprescindible trabajar en redes temáticas de referencia en las que hay modelos maduros de normalizados para la descripción de datos, tecnologías de publicación y conjuntos de datos significativos. Este conocimiento experto por disciplinas ha de ser aprovechado para diseminar en nuestro país las buenas prácticas internacionales. Los protocolos deben atacar el momento de la producción de los datos, es decir, estar orientados para ser implantados al nivel de proyecto de investigación.

Por parte de las **organizaciones** a las que pertenecen, así como en las agencias de evaluación, debe comenzar a incentivarse el compartido de datos, reconociendo como parte del trabajo científico.

Las entidades **financiadoras** deben empezar a contemplar mandatos que obliguen a depositar en abierto los datos financiados con dinero público, una vez que los protocolos y la infraestructura tecnológica esté suficientemente desarrollada.

Las vías por las que los **gestores de información** académica han de orientarse no son únicas, pero no deben duplicarse. Por una parte las organizaciones donde se producen los datos deberían implantar modelos de gestión del ciclo vital de los datos vinculados al CRIS, al repositorio…; pero la opción que se vislumbra en la actualidad con mayores ventajas y coste de gestión es el almacenamiento junto a la publicación en plataformas editoriales. La hibridación repositorio con los propios datos parece que puede ser una solución eficiente, y por la que apuesta la Unión Europea. Esta opción reduce costes de todo tipo, pero hay que tener en cuenta que en muchas ocasiones los datos a los que se refiere un artículo son solo una parte de una gran base de datos. Los países con mayor trayectoria en el almacenamiento de datos realizan el almacenamiento de manera inversa: Se almacena la base de datos y se añade la literatura vinculada con esos datos..

En todo caso, se detecta una escasez global de demanda de servicios a las bibliotecas, lo que impone una labor intensiva de promoción de nuestras capacidades entre los investigadores. Es importante el papel de los servicios bibliotecarios y la necesidad de impartir la formación apropiada al personal bibliotecario para conseguir una buena gestión de los datos (su ciclo vital en su conjunto) y una relación recíproca de confianza en el servicio entre bibliotecarios e investigadores.

Por último, los datos de investigación han de encauzarse en estos momentos en la tendencia general a la innovación propiciada por el manejo de grandes volúmenes de datos (BigData) mediante ontologías especializadas. Hay que sumarse a la tendencia a hacer más productivos los datos: "The age of data-driven science"

Los datos son cool, los datos son negocio, los datos son ciencia

11. CONCLUSIONES

El **movimiento de acceso abierto** ha provocado un debate sobre el acceso, uso y modelos de negocio de la información producida con fondos públicos, incluyendo además de las **publicaciones** científicas los **datos** de investigación.

Los datos de investigación están empezando a reconocerse como una fuente de conocimiento propia e independiente de las publicaciones que pueden emplearse en la validación de los resultados de investigación publicados en artículos, para generar nuevo conocimiento y ser explotados de manera interdisciplinar.

Para asegurar la explotación de los datos, es necesario que estén disponibles y accesibles en la red, sin embargo su **naturaleza** es **muy heterogénea**, dependiente de la disciplina y de su particular ciclo de vida. Como consecuencia los requisitos técnicos y legales para garantizar el acceso son complejos. Los datos, según su origen, se pueden categorizar en: observacionales, experimentales y computacionales.

Una **gestión** adecuada de los **datos** requiere al menos los siguientes aspectos:

- **Políticas**, a nivel de agencias de financiación e institucionales, que definan los papeles y las responsabilidades de los distintos actores.
- **Recursos financieros** a largo plazo ya que los datos son acumulativos y se preservan.
- **Recursos humanos** especializados (para generación de datos, explotación y preservación).
- **Infraestructuras** coordinadas para garantizar su interoperabilidad. Entre los requisitos de las infraestructuras destacar: preservación, acceso, data curation, data processing, distribución.
  Para dar respuesta a estos aspectos es necesaria una formación adecuada (tanto para sus creadores como para los responsables de su mantenimiento), equipamientos, sistemas de almacenamiento masivo de datos y redes de alta capacidad.
- **Cambio cultural** en los actores involucrados: investigadores/creadores de datos; Universidades y Centros de Investigación; Repositorios Institucionales; Centros de datos; Gestores de datos; Usuarios que reutilizan los datos; Agencias de financiación; Editores de publicaciones científicas.

Ya existe un acuerdo internacional para considerar la creación de una infraestructura transnacional y multidisciplinar que garantice el acceso a los datos de investigación.

Para estimular el depósito en abierto, diferentes organismos internacionales (UE, OCDE,...) han emitido **recomendaciones**, que están marcando tendencias, dirigidas a:

- Facilitar el acceso a las publicaciones científicas y datos.

- Cofinanciar infraestructuras de investigación (repositorios).
- Estimular el debate para futuras políticas al respecto.
- Estimular el debate entre los diferentes actores en la gestión de los datos científicos.

Y además, en el ámbito del VII Programa Marco, se llevan a cabo diversos **proyectos** piloto para la creación de e-infraestructuras, el depósito de artículos en repositorios (Ej: *OpenAire*) y el depósito de datos (Ej: *Open Aire Plus*).

La responsabilidad de la gestión de los datos corresponde a los investigadores, a las bibliotecas, a los servicios informáticos y a las instituciones en general. La creación de los datos corresponde a los investigadores pero la gestión del ciclo vital corresponde a los gestores de la información, es decir a bibliotecarios especializados. Las instituciones deben proporcionar el soporte técnico y organizativo facilitando un modelo de **plan de gestión de datos** que permita: encontrar y comprender los datos cuando se necesite utilizarlos, garantizar la continuidad del proyecto, evitar duplicidades, validar los resultados, compartir, potenciar la visibilidad si es depósito en abierto y el prestigio de la investigación al citar los datos.

Cabe resaltar que la gestión de los datos de investigación debe llevarse a cabo durante todo el proceso de investigación: antes de la creación de los datos, durante su creación y uso y a lo largo de su ciclo de vida.

En un plan de gestión de datos, que debería incluirse en toda propuesta de financiación, se debe tener en cuenta:

- **Organización y documentación** de los datos según los estándares. Mecanismos de **almacenamiento**, **back-up**, **seguridad** y **compartido** de datos.

  El ***'dataset'*** es una colección de datos reunidos durante la ejecución de un proyecto de investigación, constituye la base de una investigación y va asociado a una publicación científica. Los datasets se almacenan y gestionan en repositorios interoperables en red.

  El **formato** en el que se archivan los datos debería ser en abierto, permitir la indización del contenido (para su recuperación), la compresión de datos (menos espacio de almacenamiento) y en un formato estándar para la comunidad investigadora.

  La nomenclatura es importante para identificar los contenidos y también es necesario llevar un control de versiones.

  La documentación del dataset facilitada por el investigador se incluye en el registro de **metadatos**.

  Para lograr la interoperabilidad con otros sistemas de gestión de datos es prioritaria la normalización. Existen varios estándares de metadatos.

El dataset almacenado debe asociarse a un **identificador digital** único y los identificadores deben tener forma de URI. La URI es una cadena de caracteres que condensa la dirección URL (Uniform Resource Location) y el nombre URN (Uniform Resource Name) del recurso. Hay muchos sistemas diferentes, por ejemplo: PURL Uniform Resource Locator, DOI Digital Object Identifier, Accession, InChI.

- Cuestiones éticas y legales: en el **marco legal** de la gestión se deben tener en cuenta los derechos legales sobre los datos y datasets, propiedad intelectual, confidencialidad, privacidad y protección de datos (pueden ser públicos, restringidos o privados), contratos, permisos y las licencias a utilizar.

  Cualquier investigación que contenga datos de carácter personal tiene que cumplir con la ley de protección de datos.
  Las colecciones de datos y las bases de datos están protegidas por propiedad intelectual y se distinguen:
  - ✓ Derechos de autor: pertenecen a los creadores siempre que se trate de trabajos originales.
  - ✓ Derechos morales: son de carácter personal, pertenecen exclusivamente a los autores y son irrenunciables.
  - ✓ Derechos de explotación o copyright: son transferibles. Existen licencias alternativas al copyright como por ejemplo las Creative Commons en el que se especifican los términos en los que compartir y reutilizar los datos.

- Un **plan de preservación** de los datos conforme a estándares internacionales. Se debe archivar una copia final de los datos, en centros de datos especializados, y en diferentes tipos de soporte (tener en cuenta la obsolescencia del hardware y el software).

Para rentabilizar mejor los costes se sugieren dos principios básicos de **sostenibilidad**: seleccionar los datos en función de aquellos que puedan ser enriquecidos o preservados y uso de economías de escala en las infraestructuras para conseguir una transversalidad.

A continuación se hace un recorrido por las **iniciativas españolas** sobre **gestión de datos científicos**. Para ello se ha examinado literatura, jornadas y conferencias así como proyectos:

- En cuanto a la **literatura académica y profesional** se detecta una acuñación del término ***data curation***, se abordan los **beneficios y formas de compartir datos**, los **modelos de coste** para la planificación de una política de preservación de la información digital y los datos, los **beneficios y coste** de la puesta en marcha de una **infraestructura global de datos fiable y estable**, las **vías de actuación** ante la especial reticencia a compartir datos y las **bases de un proyecto de gestión** de datos a nivel autonómico basado en buenas prácticas internacionales.

- En las **jornadas y conferencias** relacionadas con la gestión de datos de investigación se resaltan los ***Webinars*** del proyecto Recolecta, la **jornada STM** sobre el estado de las bases de datos en química cuántica y la necesidad del proyecto Quixote ante la falta de un modelo de datos estandarizado en la disciplina del cálculo, los aspectos más reseñables del **data sharing**, la **integración de datos primarios y publicaciones**, la distinción entre datos de investigación producidos por los investigadores y los datos de los grandes productores, la jornada **Research Data Management** o la visión panorámica internacional de los avances en gestión de datos presentada en **OS-Repositorios**.
- Hay diversos **proyectos** relacionados con la gestión de datos como:
  - Proyecto *Wf4Ever*, sobre la evolución, intercambio y colaboración de flujos de trabajo.
  - Proyecto internacional, *agINFRA "Promoting data sharing and development of trust in agricultural sciences"*, para la implementación de una infraestructura abierta para mejorar la transferencia de resultados científicos y tecnológicos del área de la agricultura.
  - La *Red Paneuropea de Gestión de Datos Marinos y Oceánicos (SeaDataNet)* para el desarrollo de un sistema interoperable para la gestión de datos e información marina.
  - Proyecto *Science-a* para el tratamiento, visualización de los datos y soporte a la elaboración de proyectos.
  - ADMIRE *(Advanced Data Mining and Integration Research for Europe)* para extracción de información significativa, realizando minería de datos, y la integración a través de los perfiles de usuario.
  - El proyecto *PlanetData* para el establecimiento de una comunidad europea sostenible de investigadores que apoyen a las organizaciones en la publicación de sus datos en nuevas y útiles formas.
  - Proyecto *MyBigData* para la creación de una plataforma que integre nuevos métodos, técnicas y herramientas que permita la integración de fuentes de datos heterogéneas de carácter científico.
  - Proyecto *BabelData* para el desarrollo de técnicas y algoritmos para la construcción de servicios capaces de crear y utilizar ontologías y datos multilingües.
  - Proyecto *GeoBuddies* para el desarrollo de una aplicación que proporcione información y servicios geo-espaciales para los peregrinos del Camino de Santiago.

Se describe como caso de estudio en nuestro país *"**ODiSEA:** International Registry on Research Data"* que tiene el objetivo de facilitar la identificación de las fuentes de almacenamiento de datos de investigación que permitan conocer dónde deben depositar los investigadores sus datos y si existen lagunas disciplinares. Además tiene previstas otras utilidades: i) proporcionar datos básicos sobre los bancos registrado; ii) ser el germen del futuro meta-buscador *opendatascience* de conjuntos de datos; y iii) crear conocimiento de investigación sobre la gestión y producción de los datos de investigación.

ODiSEA cuenta con depósitos entre los se encuentran bancos especializados, bibliotecas de datos, repositorios y bancos de imágenes.

Permite buscar entre los depósitos que están registrados, según diversos criterios, y se continúa investigando para proporcionar resultados a medida que los datos van siendo depositados en abierto.

Este informe ha sido elaborado en el marco de **Recolecta** que es un proyecto, gestionado y coordinado por FECYT, para la creación de una red de repositorios científicos institucionales interoperables para facilitar la *"open science"* o ciencia en abierto, en cumplimiento del *artículo 37* la *Ley 14/2011, de 1 de Junio, de la Ciencia, la Tecnología y la Innovación*. Recolecta también tiene el objetivo de dotar de mayor visibilidad y servicios a los resultados de investigación y la producción científica española.

FECYT, junto con un **grupo de expertos**, ha elaborado el presente informe para dar apoyo a la gestión de los datos de investigación. Ha contado con la participación de expertos adscritos a las Universidades Carlos III (UC3M) y Complutense de Madrid (UCM), el Consejo Superior de Investigaciones Científicas (CSIC), la Universidad de Alicante (UA), el Centro de Servicios Científicos y Académicos de Cataluña (CESCA), el Instituto Juan March y la Universitat Politècnica de Catalunya (UPC) actuando esta última como coordinadora. Posteriormente se incorporó también a este grupo la Universidad Politécnica de Valencia (UPV).

12. BIBLIOGRAFÍA

- Australian Government. Department of Education, Science and Training (2007). *A proposal for an Australian National Data Service.* HTTP://WWW.PFC.ORG.AU/PUB/MAIN/DATA/TOWARDSTHEAUSTRALIANDATACOMMONS.PDF [Consulta 8/12/2012]
- Barragán, Antonio; Bermúdez, Óscar (2007). *Propuesta del Protocolo de remisión, almacenamiento y difusión de datos antárticos en España. Centro Nacional de Datos Polares y Archivo Polar.* HTTP://WWW.UIB.ES/DEPART/DFS/APL/AAC/AA/ANTARTIDA/PGCDCAE/08_CNDP/PROTOCOLODATOS.PDF [Consulta 9/12/2012]
- Bailey, Charles W., Jr. (2013). Research Data Curation Bibliography. HTTP://DIGITAL-SCHOLARSHIP.ORG/RDCB/RDCB.HTM [Consulta 15/01/2013]
- Bailey, Charles W., Jr. (2013). Digital Curation Bibliography: Preservation and Stewardship of Scholarly Works. HTTP://DIGITAL-SCHOLARSHIP.ORG/RDCB/RDCB.HTM [Consulta 15/01/2013]
- Beguería S., Vicente-Serrano S.M., Angulo M. A multi-scalar global drought data set: the SPEIbase. *Bulletin of the American Meteorological Society*, DOI: 10.1175/2010BAMS2988.1.
- Bermúdez Molina, Oscar; Barragán Sanabria, Antonio; Alonso Gallego, Francisco (2007). Evaluación de la producción científica de la investigación española en la Antártida. *10as Jornadas Españolas de Documentación: FESABID 2007* Santiago de Compostela, 9-11 de mayo.
- Bermúdez, Óscar; Barragán, Antonio; Alonso, Francisco (2011). La gestión de los datos polares en España: una aproximación a la contribución de las ciencias de la vida. *Ecosistemas*, v. 20, n.1, pp. 94-103. HTTP://REVISTAECOSISTEMAS.NET/PDFS/675.PDF [Consulta 9/12/2012]
- Borgman, C.L. (2011). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology.* HTTP://PAPERS.SSRN.COM/SOL3/PAPERS.CFM?ABSTRACT_ID=186915 [Consulta 8/12/2012]
- Borgman, C.L. (2012). *On Local or Global? Making Sense of the Data Sharing Imperative. Talk at University of Southern Carolina on 9th April 2012*
- Borrego, Angel (2012). Enriquecer las publicaciones con datos empíricos. *Reseñas de Biblioteconomía y Documentación.* ISSN: 2014-0894, HTTP://WWW.UB.EDU/BLOKDEBID/ES/CONTENT/ENRIQUECER-LAS-PUBLICACIONES-CON-DATOS-EMPÍRICOS-0 [Consulta 9/12/2012]
- Borrego, Angel (2012). Los retos de la gestión de datos de investigación. *Reseñas de Biblioteconomía y Documentación,* ISSN: 2014-0894, HTTP://WWW.UB.EDU/BLOKDEBID/ES/CONTENT/LOS-RETOS-DE-LA-GESTIÓN-DE-DATOS-DE-INVESTIGACIÓN [Consulta 9/12/2012]

- Botella Ausina, Juan; Ortego Maté, María del Carmen (2010). Compartir datos: hacia una investigación más sostenible. *Psicothema,* Vol. 22, n 2. págs. 263-269 Disponible en: HTTP://WWW.PSICOTHEMA.COM/PDF/3725.PDF [Consulta 9/12/2012]
- Castro Martín, Pablo de; García Gómez, Consol; Rodríguez Miranda, Álvaro (2012). Gestión de datos de investigación en repositorios de acceso abierto: una visión panorámica y un caso práctico en la UPV/EHU. *Jornadas Os-Repositorios (5as Bilbao, 23 al 25 de mayo). Universidad del País Vasco*
- Castro, Pablo de (2012). Avances recientes a nivel internacional en la gestión de datos de investigación. *Advances in Research Data Management (Barcelona, 10 mayo) GrandIR / Universitat Politècnica de Catalunya.* HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME [Consulta 9/12/2012]
- Christensen-Dalsgaard, Birte et al (2012*). Ten recommendations for libraries to get started with research data management. Final report of the LIBER working group on E-Science / Research Data Management.* HTTP://WWW.LIBEREUROPE.EU/SITES/DEFAULT/FILES/WGSC_20120801.PDF [Consulta 9/12/2012]
- Comisión Europea (2010). *Una agenda Digital para Europa,* HTTP://EUR-LEX.EUROPA.EU/LEXURISERV/LEXURISERV.DO?URI=COM:2010:0245:FIN:ES:PDF [Consulta 9/12/2012]
- *Commission Decision on the adoption and a modification of special clauses applicable to the model grant agreement of FP7 C(2008) 4408 final* HTTP://EC.EUROPA.EU/RESEARCH/PRESS/2008/PDF/DECISION_GRANT_AGREEMENT.PDF [Consulta 8/12/2012]
- *Communication on scientific information in the digital age: access, dissemination and preservation (Com 2007)56);* HTTP://EC.EUROPA.EU/RESEARCH/SCIENCE-SOCIETY/DOCUMENT_LIBRARY/PDF_06/COMMUNICATION-022007_EN.PDF [Consulta 8/12/2012]
- National Science Foundation(2007). *Cyberinfrastructure Vision for 21st Century Discovery* HTTP://WWW.NSF.GOV/PUBS/2007/NSF0728/INDEX.JSP [Consulta 8/12/2012]
- *Digital Curation Center. All standards for any lifecycle action.* HTTP://WWW.DCC.AC.UK/RESOURCES/STANDARDS/DIFFUSE/STANDARDS?FRAMEWORK_ID=0&LIFECYCLE_ID=0&SORT=TYPE [Consulta 8/12/2012]
- Echenique, Pablo (2011). The Quixote Project: a pioneering work in managing Computational Chemistry research data. *STM Research Data Management.* Grandir ZCAM, (Zaragoza, August 25, 2011) HTTP://DIGITAL.CSIC.ES/BITSTREAM/10261/39026/1/2011_08_QUIXOTE_MEETING.PDF [Consulta 9/12/2012]
- Estrada, Jorge; Echenique, Pablo (2011). From Databases in QC 2010, ZCAM, Sep 2010 onwards: a brief history of Quixote. *STM Research Data Management.* Grandir ZCAM, (Zaragoza, August 25, 2011) HTTP://DIGITAL.CSIC.ES/BITSTREAM/10261/39038/1/2011_8_25_QUIXOTE_MEETING_V2.PDF [Consulta 9/12/2012]
- Estrada, Marta; Álvarez, Enrique; Barragán, Antonio; Bermúdez, Óscar; García, Mª Jesús; Lavín, Alicia; Masqué, Pere; Pérez, Fiz F; Piera, Jaume (2011). *INFORME SCOR Comité Científico sobre*

Investigación Oceánica Representación española. Reflexiones sobre la gestión y la custodia de datos oceanográficos en España. Recursos existentes y recomendaciones para el futuro. HTTP://WWW.SCOR-ES.ORG/DOCUMENTACION/REFLEXIONES_GESTION_DATOS.PDF [Consulta 9/12/2012]

- European Commission. (2010). *Global Research Data Infrastructures: The GRDI2020 Vision. GRDI2020 project.* HTTP://WWW.GRDI2020.EU/REPOSITORY/FILESCARICATI/FC14B1F7-B8A3-41F8-9E1E-FD803D28BA76.PDF [Consulta 8/12/2012]
- European Union. (2007). *Council Conclusions on scientific information in the digital age: access, dissemintation and preservation;* HTTP://WWW.CONSILIUM.EUROPA.EU/UEDOCS/CMS_DATA/DOCS/PRESSDATA/EN/INTM/97236.PDF [Consulta 8/12/2012]
- Ferrer-Sapena, A.; Villamón, M.; González-Moreno, L.M.; Peset, F.; Aleixandre, R.; García-García, A.; Morales-Aznar, J. Gestión de los datos de investigación como medida de calidad. Mª Teresa Ramiro, Mª Paz Bermúdez e Inmaculada Teva (Comps.) (2012). *Evaluación de la Calidad de la Investigación y de la Educación Superior (IX Foro)*, pp. 483. Granada: Asociación Española de Psicología Conductual (AEPC). ISBN: 978-84-695-3701-5.
- García-García, A.; García-Massó, X.; Ferrer, A.; González-Moreno, L.M.; Peset, F.; Aleixandre, R. Mejores prácticas en reúso de conjuntos de datos publicados online como material adicional a los artículos. *2a Conferencia sobre calidad de revistas de ciencias sociales y humanidades* (CRECS 2012) HTTP://WWW.THINKEPI.NET/CRECS2012 [Consulta 9/12/2012]
- García-García, Alicia; García–Massó, Xavi; Ferrer-Sapena, Antonia; González, Luis-Millán; Peset, Fernanda; Rodríguez-Gairín, Josep-Manuel; Saorín, Tomás (2012). ODiSEA: International Registry on Research Data. *5as Jornadas OS-Repositorios "La motricidad de los repositorios de acceso abierto"* 23 al 25 de mayo 2012. Universidad de País Vasco.
- Greenberg, Jane (2009). Theoretical Considerations of Lifecycle Modeling: An Analysis of the Dryad Repository Demonstrating Automatic Metadata Propagation, Inheritance, and Value System Adoption. *Cataloging & Classification Quarterly*, vol. 47, núm. 3, p. 380-402.
- *Is it Open Data?* HTTP://ISITOPENDATA.ORG/ [Consulta 8/12/2012]
- Jefatura del Estado. (2011). Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación. *Boletín Oficial del Estado,* vol. núm. 131, no. 2 de junio de 2011, pp. 54387 a 54455. HTTP://WWW.BOE.ES/BOE/DIAS/2011/06/02/PDFS/BOE-A-2011-9617.PDF [Consulta 9/12/2012]
- Keefer, Alice (2011). La preservación de los datos de investigación y las agencias de financiación de la I+D. *Reseñas de Biblioteconomía y Documentación.* ISSN: 2014-0894, HTTP://WWW.UB.EDU/BLOKDEBID/ES/NODE/130 [Consulta 5/1/2013]
- Lacunza, Izaskun**.** OpenAIREplus: a European initiative as a driver for national RDM activity. *Advances in Research Data Management* (Barcelona, 10 mayo 2012) GrandIR / Universitat Politècnica de Catalunya HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME [Consulta 9/12/2012]

- Lahoz, José María (2012). Una perspectiva de la gestión de datos desde las Humanidades. *Advances in Research Data Management* (*Barcelona, 10 mayo) GrandIR / Universitat Politècnica de Catalunya.* HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME [Consulta 9/12/2012]
- Lyon, Liz (2007). Dealing with Data: Roles, Rights, Responsabilities and Relationships. Consultancy Report. UKOLN. HTTP://WWW.UKOLN.AC.UK/UKOLN/STAFF/E.J.LYON/REPORTS/DEALING_WITH_DATA_REPORT-FINAL.DOC [Consulta 8/12/2012]
- Lyon, Liz (2012). The Informatics Transform: Re-Engineering Libraries for the Data Decade. *The International Journal of Digital Curation.* Volume 7, Issue 1, 2012. HTTP://WWW.IJDC.NET/INDEX.PHP/IJDC/ARTICLE/VIEW/210/279 [Consulta 8/12/2012]
- Managing and sharing data. Best practice for researchers. *UK Data Archive,* 2011 (rev.). HTTP://WWW.DATA-ARCHIVE.AC.UK/MEDIA/2894/MANAGINGSHARING.PDF [Consulta 8/12/2012]
- Marcos-Martín, Carlos; Soriano-Maldonado, Salvador-Luis (2011). Reutilización de la información del sector público y Open data en el contexto español y europeo. Proyecto Aporta. *El profesional de la información,* vol. 20, núm. 3. p.291-297 HTTP://ADMINISTRACIONELECTRONICA.GOB.ES/RECURSOS/PAE_020002228.PDF [Consulta 9/12/2012]
- Martínez-Uribe, Luis, Macdonald, Stuart (2008). Un nuevo cometido para los bibliotecarios académicos: data curation. *El profesional de la información*, v.17, n. 3, mayo-junio 2008
- Martinez-Uribe, Luis; Fernández, Paz (2011). La Biblioteca de Datos del Centro de Estudios Avanzados en Ciencias Sociales (CEACS) del Instituto Juan March como un servicio de apoyo a su comunidad científica. *Webinar FECYT/Recolecta sobre almacenamiento, conservación y gestión de los datos de investigación.* 7 de noviembre al 19 de diciembre. Disponible en: HTTP://WWW.RECOLECTA.NET/BUSCADOR/WEBMINARS_PDF/CEACS_DATA_LIBRARY.PDF [Consulta 9/12/2012]
- Martínez-Uribe, Luis; Macdonald, Stuart (2008). Un nuevo cometido para los bibliotecarios académicos: data curation. *El profesional de la información*, vol. 17, núm. 3, p. 273-280. HTTP://WWW.ELPROFESIONALDELAINFORMACION.COM/CONTENIDOS/2008/MAYO/03.PDF [Consulta 9/12/2012]
- Martinez-Uribe, Luis; Macdonald, Stuart (2009). User Engagement in Research Data Curation. *Lecture Notes in Computer Science,* vol. 5714, p. 309-314.
- Matorras, Francisco (2009). The CMS Computing Model, HTTP://INDICO.CERN.CH/GETFILE.PY/ACCESS?CONTRIBID=2&RESID=0&MATERIALID=SLIDES&CONFID=68690 [Consulta 9/12/2012]
- Melero, Remedios (2010). Una pleamar de datos. *Reseñas de Biblioteconomía y Documentación.* ISSN: 2014-0894, HTTP://WWW.UB.EDU/BLOKDEBID/ES/CONTENT/UNA-PLEAMAR-DE-DATOS [Consulta 9/12/2012]

- Murillo, Angela; Greenberg, Jane (2012). Data-at-Risk, Metadata Registration for Data, and Dryad. *Advances in Research Data Management* (Barcelona, 10 mayo 2012) GrandIR / Universitat Politècnica de Catalunya HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME [Consulta 9/12/2012]
- OECD (2007). *Principles and Guidelines for Access to Research Data from Public Funding;* HTTP://WWW.OECD.ORG/DATAOECD/9/61/38500813.PDF [Consulta 8/12/2012]
- Parlamento Europeo (2003). Directiva 2003/98/ce del Parlamento Europeo y del Consejo de 17 de noviembre de 2003 relativa a la reutilización de la información del sector público. *Diario Oficial de la Unión Europea,* vol. Num. 345, no. 31 de diciembre de 2003 HTTP://EU.VLEX.COM/SOURCE/DOUE-23/ISSUE/2003/12/31/1 [Consulta 9/12/2012]
- Payne, Geoff; Treloar, Andrew (2006). The ARROW Project after two years: are we hitting our targets?. *Proceedings of VALA*, Melbourne. HTTP://WWW.VALACONF.ORG.AU/VALA2006/PAPERS2006/57_TRELOAR_FINAL.PDF [Consulta 9/12/2012]
- Pérez González, Lourdes (2010). Modelo/s de coste para la preservación de los datos científicos en la e-ciencia. *XII Jornadas de Gestión de la Información. Valor económico de la información: mercados, servicios y rentabilidad.* SEDIC, 18-19 de noviembre. Disponible en: HTTP://EPRINTS.RCLIS.ORG/BITSTREAM/10760/8555/1/PEREZ.PDF [Consulta 9/12/2012]
- Pérez González, Lourdes (2011). E-ciencia y la información como bien público, algunas propuestas. *XIII Jornadas de Gestión de la Información.* BNE, Madrid 17 y 18 nov. Disponible en: HTTP://WWW.SEDIC.ES/SERVICIOS-ETICA-DCHOS-HUMANOS.PDF [Consulta 9/12/2012]
- Pérez González, Lourdes (2011). Towards a Galician Data Commons. 75th Annual Meeting of the Society of American Archivists. HTTP://DLC.DLIB.INDIANA.EDU/DLC/BITSTREAM/HANDLE/10535/7870/TOWARDS%20A%20GALICIAN%20DATA%20COMMONS.PDF?SEQUENCE=1 [Consulta 9/12/2012]
- Pérez, Esther; Maestre, Roberto; Bosque, Isabel del; Crespo Solana, Ana; Sánchez-Crespo, Juan Manuel (2010). DynCoopNet-CSIC-Objetivos y Estado Actual del Proyecto HTTP://HUMANIDADES.CCHS.CSIC.ES/CCHS/SIG/PDF/PDF/DYNCOOPNET/ESTHERPEREZ_DYNCOOPNET.PDF [Consulta 9/12/2012]
- Pérez, Esther; Maestre, Roberto; Bosque, Isabel del; Crespo Solana, Ana; Sánchez-Crespo, Juan Manuel (2012). Modelling and Implementation of a spatio-temporal historic GIS. Self-organizing Networks and GIS Tools. Cases of Use for the Study of Trading Cooperation (1400-1800). *Journal of Knowledge Management, Economics and Information Technology.* HTTP://DIGITAL.CSIC.ES/BITSTREAM/10261/59170/1/MODELLING_AND_IMPLEMENTATION_OF_A_SPATIOTEMPORAL_HISTORIC_GIS.PDF [Consulta 9/12/2012]
- Pérez, Esther; Maestre, Roberto; Bosque, Isabel del; Crespo Solana, Ana; Sánchez-Crespo, Juan Manuel (2010). Integración de bases de datos históricas en una IDE. Comercio mundial y redes de cooperación en la primera Edad Global (1400-1800), *Jornadas Técnicas de la Infraestructura de Datos Espaciales de España*

HTTP://DIGITAL.CSIC.ES/BITSTREAM/10261/24908/1/JIDEE09_DYNCOOPNET_FINAL.PDF [Consulta 9/12/2012]

- Peset, F.; Aleixandre, R.; Villamón, M.; González-Moreno, L.M.; Ferrer, A. (2012). Open Data en el mundo científico: proyecto OpenDataScience. En: Lidia Cabello y Mª Paz Bermúdez (Comps.) (2011). *Evaluación de la Calidad de la Investigación y de la Educación Superior (VIII Foro FECIES)*, pp. 481-482. Granada: Asociación Española de Psicología Conductual (AEPC). ISBN: 978-84-694-3488-8.
- Peset, Fernanda (2012). Opiniones del sector científico sobre la preservación de la información. Blok de BiD. *Reseñas de Biblioteconomía y Documentación.* ISSN: 2014-0894, septiembre. HTTP://WWW.UB.EDU/BLOKDEBID/ES/CONTENT/OPINIONES-DEL-SECTOR-CIENT%C3%ADFICO-SOBRE-LA-PRESERVACIÓN-DE-LA-INFORMACIÓN [Consulta 9/12/2012]
- *Riding the Wave: How Europe can gain from the rising tide of scientific data* (2010). HTTP://CORDIS.EUROPA.EU/FP7/ICT/E-INFRASTRUCTURE/DOCS/HLG-SDI-REPORT.PDF [Consulta 8/12/2012]
- Serrano-Muñoz, Jordi (2012). FECYT/Recolecta Working Group for Data Repositories. Advances in Research Data Management (Barcelona, 10 mayo) GrandIR / Universitat Politècnica de Catalunya. HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME [Consulta 9/12/2012]
- Sorribas Cervantes, Jordi (2012). La gestión de datos (marinos) desde la perspectiva de un centro de datos de investigación. *Advances in Research Data Management* (Barcelona, 10 mayo) GrandIR / Universitat Politècnica de Catalunya. HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME [Consulta 9/12/2012]
- Special Online Collection: Dealing with Data (2011). *Science.* 11 February HTTP://WWW.SCIENCEMAG.ORG/SITE/SPECIAL/DATA/ [Consulta 8/12/2012]
- Torres-Salinas, Daniel (2010). Compartir datos (data sharing) en ciencia: contexto de una oportunidad. *Anuario ThinkEPI*, vol. 4, p. 258-261. HTTP://WWW.THINKEPI.NET/COMPARTIR-DATOS-DATA-SHARING-EN-CIENCIA-EL-CONTEXTO-DE-UNA-OPORTUNIDAD [Consulta 9/12/2012]
- Torres-Salinas, Daniel (2010). Hacia la gestión de datos de investigación en las universidades: la Data asset framework. *Anuario ThinkEPI*, vol.4, p. 262-265. HTTP://WWW.THINKEPI.NET/PRIMEROS-PASOS-HACIA-LA-GESTION-DE-DATOS-DE-INVESTIGACION-EN-LAS-UNIVERSIDADES-LA-INICIATIVA-DAF [Consulta 9/12/2012]
- Torres-Salinas, Daniel; Robinson-García, Nicolás; Cabezas-Clavijo, Álvaro (2012). Compartir los datos de investigación: introducción al data sharing. *El profesional de la información*, marzo-abril, vol. 21, n. 2, p. 173-184. HTTP://HDL.HANDLE.NET/10760/16786 [Consulta 9/12/2012]
- Treloar, A (2006). The Dataset Acquisition, Accessibility, and Annotation e-Research Technologies (DART) Project: building the new collaborative e-research infrastructure. *Proceedings of AusWeb06, the Twelfth Australian World Wide Web Conference*, Southern Cross University Press, Southern

Cross University, July. HTTP://AUSWEB.SCU.EDU.AU/AW06/PAPERS/REFEREED/TRELOAR/PAPER.HTML [Consulta 9/12/2012]

- Treloar, A.; Groenewegen, D.; Harboe-Ree, C. (2007). The Data Curation Continuum. Managing Data Objects in Institutional Repositories. *D-Lib Magazine*, vol. 13, núm 9/10. HTTP://WWW.DLIB.ORG/DLIB/SEPTEMBER07/TRELOAR/09TRELOAR.HTML [Consulta 9/12/2012]
- *University of Melbourne Research Data Management Policy* HTTP://RESEARCH.UNIMELB.EDU.AU/INTEGRITY/CONDUCT/DATA/REVIEW [Consulta 8/12/2012]
- Vallverdù, Francesc (2012). Research Data Management: A Perspective from a University Department. *Advances in Research Data Management* (Barcelona, 10 mayo) GrandIR / Universitat Politècnica de Catalunya. HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME [Consulta 9/12/2012]
- Van der Graaf, M. and Waaijers, L. (2011). *A Surfboard for Riding the Wave. Towards a four country action programme on research data.* HTTP://WWW.KNOWLEDGE-EXCHANGE.INFO/SURFBOARD [Consulta 8/12/2012]
- Vicente-Serrano S.M., Beguería S., López-Moreno J.I. (2010). A Multi-scalar drought index sensitive to global warming: The Standardized Precipitation Evapotranspiration Index – SPEI. *Journal of Climate* 23(7), 1696-1718, DOI: 10.1175/2009JCLI2909.1
- Vicente-Serrano S.M., Beguería S., López-Moreno J.I., Angulo M., El Kenawy A. A global 0.5° gridded dataset (1901-2006) of a multiscalar drought index considering the joint effects of precipitation and temperature. *Journal of Hydrometeorology* 11(4), 1033-1043, DOI: 10.1175/2010JHM1224.1.
- Wacowicz, Monica; Crespo Solana, Ana; Bernabé Poveda, Miguel Ángel (2010). Visualization and Space Time representation of Dynamis, non linear Spatial Data in DynCoopNet Project. HTTP://DIGITAL.CSIC.ES/BITSTREAM/10261/23414/1/SCIENTIFICREPORT.%20WACHOWICZ.CRESPO.BERNABE.PDF [Consulta 9/12/2012]
- Zúñiga, Anna (2012). Reptes i dificultats en la implementació d'estratègies institucionals per la gestió de dades. *Advances in Research Data Management* (Barcelona, 10 mayo) GrandIR / Universitat Politècnica de Catalunya. HTTP://WWW.GRANDIR.COM/EN/TECNICAL-SESSION/ADVANCES-IN-RESEARCH-DATA-MANAGEMENT-IN-SPAIN/PROGRAMME [Consulta 9/12/2012]

Sobre las instituciones participantes

La **Fundación Española para la Ciencia y la Tecnología (FECYT)** es una fundación pública, dependiente del Ministerio de Economía y Competitividad que, bajo los principios de racionalización, transparencia y eficiencia, trabaja para desarrollar instrumentos de participación social a favor de la ciencia; ser una herramienta adecuada para la divulgación de la ciencia y el incremento de la cultura científica; transformarse en el canal de comunicación con la comunidad de científicos españoles en el exterior y convertirse en un referente métrico de la I+D+i española. FECYT además apoya las estructuras de gestión de la información y recursos científicos.

Entre las actividades de FECYT se lleva a cabo el proyecto RECOLECTA[40] que coordina la creación de una red de repositorios institucionales interoperables y puede ser considerado como la primera iniciativa nacional en la creación de una infraestructura que facilita la "*open science*" o ciencia en abierto. El objetivo es además dotar de mayor visibilidad y servicios a los resultados de la investigación y de la producción científica española.

La **Universidad Carlos III (UC3M)**[41] de Madrid fue creada en 1989, y en 2010 obtuvo la acreditación de Campus de Excelencia Internacional. La Universidad consta de tres centros: la Facultad de Ciencias Sociales y Jurídicas, la Facultad de Humanidades y la Escuela Politécnica Superior ubicados físicamente en tres campus diferentes en Getafe, Leganés y Colmenarejo.

La **Universidad Complutense de Madrid (UCM)**[42] fue fundada en Alcalá de Henares por el Cardenal Cisneros en 1499. En 2009 obtuvo la acreditación como Campus de Excelencia Internacional. La Universidad Complutense cuenta con dos campus: el de Moncloa y el de Somosaguas. Sus 78 títulos de grado, que cubren una amplia gama de especialidades, se agrupan en cinco ramas de conocimiento: Humanidades, Ciencias Experimentales, Ciencias de la Salud, Ciencias Sociales y Jurídicas, y Tecnología.

La **Agencia Estatal Consejo Superior de Investigaciones Científicas (CSIC)**[43] es la mayor institución pública dedicada a la investigación en España y la tercera de Europa. Adscrita al Ministerio de Economía y Competitividad[44], a través de la Secretaría de Estado de Investigación, Desarrollo e Innovación, su objetivo fundamental es desarrollar y promover investigaciones en beneficio del progreso científico y tecnológico, para lo cual está abierta a la colaboración con entidades españolas y extranjeras. El motor de la investigación lo forman sus centros e institutos, distribuidos por todas las comunidades autónomas, y sus más de 15.000 trabajadores, de los cuales más de 3.000 son investigadores en plantilla y otros tantos

[40] http://www.recolecta.net [Consulta 6/12/2012]
[41] http://www.uc3m.es [Consulta 6/12/2012]
[42] http://www.ucm.es/ [Consulta 6/12/2012]
[43] http://www.csic.es/ [Consulta 6/12/2012]
[44] http://www.micinn.es/ [Consulta 6/12/2012]

doctores y científicos en formación. Por su carácter multidisciplinar y multisectorial el CSIC cubre todos los campos del conocimiento. Su actividad, que abarca desde la investigación básica hasta el desarrollo tecnológico, se organiza en torno a ocho áreas científico-técnicas. Además, el CSIC gestiona un conjunto de importantes infraestructuras, la red más completa y extensa de bibliotecas especializadas y cuenta con unidades mixtas de investigación. Como resultado de la firma de la Declaración de Berlín en 2006, el repositorio institucional DIGITAL.CSIC nació en 2008 y en el 2010 los datos puros se incorporaron como tipología de contenidos. La experiencia pionera en este sentido fue *SPEIbase: a global 0.5º gridded SPEI data base,* protagonista del primer número del Boletín CSIC Abierto[45].

La **Universidad de Alicante (UA)**[46] fue creada en octubre de 1979 sobre la estructura del Centro de Estudios Universitarios (CEU), que había comenzado a funcionar en 1968. La Universidad cuenta con una cincuentena de titulaciones, más de sesenta Departamentos Universitarios y unidades y grupos de investigación en Áreas de Ciencias Sociales y Jurídicas, Experimentales, Tecnológicas, Humanidades, Educación y Ciencias de la Salud, así como quince Institutos Universitarios e Interuniversitarios de investigación y nueve sedes universitarias.

La **Universitat Politècnica de Catalunya BarcelonaTech (UPC)**[47], fue creada en 1971 y está especializada en los ámbitos de la ingeniería, la arquitectura y las ciencias. Con presencia en ocho ciudades catalanas, la UPC se siente cercana a su entorno y parte activa de su desarrollo económico, cultural y social. Con esta voluntad de proximidad y de servicio, la UPC también tiene presencia en todo el mundo.

La **Universitat Politècnica de València (UPV)**[48] , con rango universitario desde 1971, es una institución pública, dinámica e innovadora, dedicada a la investigación y a la docencia que, al mismo tiempo que mantiene fuertes vínculos con el entorno social en el que desarrolla sus actividades, opta por una decidida presencia en el extranjero. Uno de los pilares del reconocimiento social de la Universitat Politècnica de València ha sido y es su capacidad investigadora. Sus departamentos, centros de investigación e institutos realizan proyectos de investigación aplicada conjuntamente con entidades y empresas nacionales e internacionales.

El **Centre de Serveis Científics i Acadèmics de Catalunya (CESCA)**[49] gestiona infraestructuras basadas en las tecnologías de la información y la comunicación (e-infraestructuras) para dar servicio a la universidad y a la investigación. El Centro tiene la visión de ser líderes en la gestión y el uso de las TIC para mejorar la calidad y la eficiencia del sistema universitario y de investigación, aprovechando las economías de escala

[45] http://digital.csic.es/handle/10261/26261 [Consulta 6/12/2012]
[46] http://www.ua.es/ [Consulta 6/12/2012]
[47] http://www.upc.edu [Consulta 6/12/2012]
[48] http://www.upv.es [Consulta 13/12/2012]
[49] http://www.cesca.cat/ [Consulta 13/12/2012]

mediante la cooperación interuniversitaria, las buenas prácticas profesionales, y la compartición de recursos.

La **Fundación Juan March** [50] fue creada en 1955. En 1987 nace el Instituto Juan March de Estudios e Investigaciones y dependiendo de éste, el Centro de Estudios Avanzados en Ciencias Sociales (CEACS). En la actualidad el CEACS es un centro de investigación postdoctoral que apoya la investigación de sus investigadores contratados y de la comunidad científica del Centro en su conjunto. Desde 1991 viene comprando bases de datos cuantitativas a organismos internacionales y a proveedores de datos y a partir de 2010 cuenta con una Biblioteca de datos y un Repositorio de datos científicos en Ciencias Sociales alojado en la Universidad de Harvard.

[50] http://www.march.es/ [Consulta 13/12/2012]


GOBIERNO
DE ESPAÑA
MINISTERIO
DE ECONOMÍA
Y COMPETITIVIDAD
FECYT
FUNDACIÓN ESPAÑOLA
PARA LA CIENCIA
Y LA TECNOLOGÍA